Translationese and Register Variation in English-To-Russian Professional Translation

Maria Kunilovskaya and Gloria Corpas Pastor

Abstract This study explores the impact of register on the properties of translations. We compare sources, translations and non-translated reference texts to describe the linguistic specificity of translations common and unique between four registers. Our approach includes bottom-up identification of translationese effects that can be used to define translations in relation to contrastive properties of each register. The analysis is based on an extended set of features that reflect morphological, syntactic and text-level characteristics of translations. We also experiment with lexis-based features from n-gram language models estimated on large bodies of originally- authored texts from the included registers. Our parallel corpora are built from published English-to-Russian professional translations of general domain mass-media texts, popular-scientific books, fiction and analytical texts on political and economic news. The number of observations and the data sizes for parallel and reference components are comparable within each register and range from 166 (fiction) to 525 (media) text pairs; from 300,000 to 1 million tokens. Methodologically, the research relies on a series of supervised and unsupervised machine learning techniques, including those that facilitate visual data exploration. We learn a number of text classification models and study their performance to assess our hypotheses. Further on, we analyse the usefulness of the features for these classifications to detect the best translationese indicators in each register. The multivariate analysis via text classification is complemented by univariate statistical analysis which helps to explain the observed deviation of translated registers through a number of translationese effects and detect the features that contribute to them. Our results demonstrate that each register generates a unique form of translationese that can be only partially explained by cross-linguistic factors. Translated registers differ in the amount and type of prevalent translationese. The

M. Kunilovskaya (✉) · G. Corpas Pastor
Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK
e-mail: maria.kunilovskaya@wlv.ac.uk

G. Corpas Pastor
University of Malaga, Malaga, Spain

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_6

same translationese tendencies in different registers are manifested through different features. In particular, the notorious shining-through effect is more noticeable in general media texts and news commentary and is less prominent in fiction.

Keywords Parallel corpora · Register variation · Translationese trends · Translationese indicators · Machine learning

1 Motivation and Aim

In this chapter we explore and compare translationese effects across several registers in English-to-Russian translation. This research builds on the long-established assumption that the intralinguistic variation between registers can be greater than the cross-linguistic differences between the same registers, famously demonstrated by Biber (1999). We also assume that the cross-linguistic differences are one of the major factors that shape the linguistic make-up of translations. The configuration of differences and similarities between the source language (SL) and the target language (TL) creates a unique language gap in each register and underlies the shining-through effect (Teich 2003) or interference, i.e. the tendency of translated texts to follow the SL patterns rather than conform to the regularities of the TL. Based on these assumptions, we are interested in establishing how the cross-linguistic distance between registers plays out with respect to the properties of translated texts in these registers.

It is especially interesting because the features used in this research to distinguish translations from the originally-authored texts in the target language (also referred to as non-translations or reference texts) are partly inspired by the variational linguistics studies that compare registers (Biber 1988; Katinskaya and Sharoff 2015; Neumann 2013; Nini 2015).

Besides variational studies, our feature selection and engineering process were guided by the previous translationese studies and evidence from the empirical translation studies, especially those that relied on interpretable (rather than surface) linguistic features to describe the typical deviations from TL norm observed in translations. Briefly, we use two feature sets: (i) frequencies of a number of morphosyntactic categories extracted from Universal Dependencies (UD) annotations and (ii) lexical frequency features that reflect the differences in the distribution of n-grams in translated and non-translated language (a detailed description of features is offered in Sect. 3.1; the description of the morphosyntactic features is offered in Appendix).

Typical translationese features for English-to-Russian translation include the overuse of relative clauses, copula verbs, modal predicates, analytical passives, generic nouns and all types of pronouns as shown below. Probably, none of the translation in the examples can be considered ungrammatical in Russian, but there is a Master Yoda-style foreign sound to them. Note that the back translations may

come across as perfectly acceptable sentences, because the translations are very literal in the first place. All examples are real-life student translations from Russian Learner Translator corpus (Kutuzov and Kunilovskaya 2014).[1]

(1) Necklaces, at first as pectorals that covered the whole chest, evolved from the prehistoric pendants. Ожерелье—первое нагрудное украшение, **которое** занимало место на всей груди, **которое** стало основой для подвесок [Necklace—first chest decoration**, which** covered the whole chest**, which** became the basis for pendants].

(2) …there are many self-employed people who manage to get money from others by means of falsely pretending to provide them with some benefit or service… Более того, **есть** много **людей**, работающих на себя, **которые** получают деньги обманным путем [Moreover, many **people are**, working for themselves**, who** get the money in a deceitful way].

(3) …differences in self-efficacy may simply mean that some teachers struggle to identify solutions to problems beyond their circle of control. …разница в самооценке **может означать** лишь **то, что** некоторые учителя испытывают сложности в нахождении решений задач **за пределами того, чем** они **могут управлять** […difference in self-evaluation **can** mean only **that** some teachers run into difficulties in finding solutions to tasks beyond the scope of **that what** they **can** control].

(4) It was difficult and exhausting to see. **Это было** тяжело и утомляюще пытаться видеть. [**It was** hard and exhausting to try to see].

These examples demonstrate a number of translation solutions that explain the increase in the frequency of TL items that are less frequent in non-translated TL than their literal counterparts in the SL. In example (2) the generic noun 'people' is rendered with a less frequent literal 'люди', instead of using a structure with zero subject or other more acceptable ways of expressing unspecified subjects. English and Russian have contrastive ways of expressing subjective modality: modal verbs are a less common choice in non-translated Russian, which prefers parenthetical means of expressing modality. The translation solution in (3) carries over the typical English modal predicate. Example (4) has the notorious literal renderings of the structures with the introductory it, which contributes to the boost of pronouns and copula verbs in translated Russian. Besides, such renditions have a strange word order, which usually interferes with the smooth flow of information in the text. Another source of surplus function words, including pronouns is the tendency to unpack the information from various concise English structures using strings of relative clauses, instead of repackaging the information in a more natural way (see (1) and (3)). Finally, example (2) demonstrates the tendency towards the explicit use of copula verbs in contexts, where a zero copula is typical in Russian.

The overarching goal of this research is to reveal and describe the register-related specificity of English-to-Russian translations in four registers.

[1] https://www.rus-ltc.org/search.

To achieve this goal, we complete several steps and answer the following research questions:

1. How clear are the register distinctions between the translated registers compared to non-translations for the two feature sets tested, provided that the suggested features reliably distinguish registers in originally-authored Russian? If the register distinctions are diluted in translations, the standardisation hypothesis stands.
2. Do registers share translationese indicators, i.e. are there translationese indicators that cut across all registers, provided that we are able to distinguish between translations and non-translations using our features?
3. What are the most important translationese indicators and most prominent translationese trends based on the results of multivariate and univariate analyses in each register?
4. Do the top translationese indicators intersect with the major cross-linguistic differences between the same registers in English and Russian to demonstrate that interference is the most important translationese effect?

These research questions are relevant to the development of the translationese theories and methodologies. The robustness of translationese indicators across registers has to be considered while building translationese detection applications. The register-induced specificity of translations has to be taken into consideration in any translation quality estimation systems based on translationese features.

In what follows, we discuss the theoretical implications of the previous translationese and variational linguistics studies for the current research and define our key concepts (Sect. 2). Section 3describes our research data and the linguistic resources used for language modelling; it also has the description of our methods and experimental setup, starting with the feature sets. The results as per the research questions are presented and commented in Sect. 4, which is followed by their interpretation in Sect. 5. Section 6 summarises the research and outlines future work.

2 Theoretical Background

2.1 Key Concepts and Approaches

The theoretical underpinnings for this research come from translationese studies, a research direction that investigates the peculiarities of translated texts that distinguish them from non-translations. This research field is related to the tasks of testing translationese universals, translationese detection, translation direction detections (including SL identification both for human and machine translation (MT)) as well as more recent studies of translationese variation along a number of dimensions such as translation competence, quality, direction, method, etc. In our

necessarily sketchy discussion of the developments in this well-established research area below, we highlight the aspects that are most relevant for the current project.

What is 'translationese'. The foundations of this type of studies were laid by Gellerstam (1986), to whom they attribute the introduction of the term 'translationese'. Gellerstam has demonstrated that there were significant statistical differences in the frequencies of loan words and colloquialisms, among other lexical features, between translated and non-translated Swedish texts. Originally, the term was used to denote statistical deviations of the translated language from the expected target language norm manifested in a reference corpus. Diana Santos (1995) extended the lexical translationese findings to include morphological phenomena such as diverging frequencies of tense and aspect forms in English and Portuguese. Her research was based on a small bidirectional parallel corpus, which provided enough occurrences of the targeted grammatical items for manual analysis. Importantly, her research design gave access to the source text and helped to link the unusual frequencies of grammatical items to the influence of the source text. We will highlight that her understanding of translationese was limited to 'the influence of properties of the source language in a translated text in a target language' (Santos 1995: 61). Her work is relevant for this research because it explicitly mentions the impact of the distance between the languages on the properties of translations. In particular, the author hypothesises that the closer the languages, the more probability of translationese due to the ease of levelling-out the differences between them.

The term translationese is sometimes used metonymically to denote any translated material (see, Nikolaev et al. 2020; Stymne 2017, for example) or to refer to the specificity of translations induced by the SL in opposition to SL/TL-independent properties of translations known as translation universals (see Rabadán et al. 2009; Santos 1995). For the purposes of this project, translationese is defined as a property of being a translation, based on the *statistical differences in frequencies of language items between translations and non-translations in the TL regardless of their hypothesised cause, which mark translations as its own language variety*.

Main translationese effects: Shining-through and independent translationese. Important developments in the descriptive approach to translations are associated with Gideon Toury's *laws of translation* (1995) and Mona Baker's *translation universals hypotheses* (1993). To put it briefly, the former generalised the observations on the properties of translations as two major laws: the law of increasing standardisation, and the law of interference from the source text. Mona Baker's theory suggested that there are universal tendencies in translation that are independent of the source and target languages. Baker's famous definition of the universal features of translation runs as follows: 'features which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems' (Baker 1993: 243). Her initial set of hypothesised universals (among the most-tested items) included *explicitation*, i.e. the tendency to spell things out rather than leave them implicit; *simplification*, i.e. the tendency to disambiguate and to avoid any risks of misunderstanding by making

texts simpler lexically and structurally; *conventionalisation* (also known as standardisation or levelling-out), i.e. the tendency for translations to exhibit relatively higher level of homogeneity than their sources; *normalisation*, i.e. the tendency to exaggerate features of the TL and to conform to its typical patterns.

The subsequent empiric research into translation universals did not corroborate the initial 'universal' claims for the proposed hypotheses. The results on a variety of translated domains, registers, language pairs and translation varieties were mixed and contradictory. To give some examples, Corpas Pastor et al. (2008) confirmed simplification for some features associated with this trend, but not for the others. Kruger and van Rooy reported limited support for the 'more explicit, more conservative, and simplified language use in the translation corpus' (Kruger and van Rooy 2010: 26).

This is not surprising for three major reasons: (1) the mapping of particular features into descriptive translationese trends can be a matter of debate (as stated in Zanettin 2013: 25); (2) there can be differences in the extraction procedures; (3) translations from different SLs and in different registers produce diverging translationese patterns. To demonstrate some of these factors consider the findings about connectives (also referred to as discourse markers, cohesive markers or conjunctions). Corpas Pastor et al. (2008) expected fewer discourse markers in translations of medical and technical texts from English into Spanish as a sign of simplification, and indeed found that 'non-translated texts use discourse markers significantly more often' in two out of three corpus pairs (Corpas Pastor et al. 2008: 24). At the same time, Koppel and Ordan (2011), while testing on English translations of addresses given in the European Parliament (Europarl) in five other languages, reported that discourse markers were significantly more frequent in translations than in the originally-authored English texts. They were inclined to interpret it as an indication of explicitation. Generally, the increase in the frequencies of discourse markers in translated language and higher cohesiveness of translations is a relatively well-explored translationese phenomenon. However, its interpretation as a manifestation of explicitation, normalisation or SL interference varies across language pairs and text categories or is unclear in some experimental setups (Castagnoli 2009; Kunilovskaya 2017; Olohan 2001). It is especially confusing if connectives are treated individually rather than cumulatively. In Jiang and Tao (2017) the frequencies of individual discourse markers were traced to the corresponding SL items to demonstrate that they contribute to several translation universals. Similarly, Becher insisted that 'every explicitating and implicitating shift has a distinct cause' and needs to be treated on a case-to-case basis (Becher 2011: 215).

In this research we refrain from assigning individual features (indicators) to the trends such as simplification and explicitation a priori. Instead, we follow a bottom-up approach and identify the indicators of some *translationese effects* based on the similarity of their frequency pattern in the source texts (ST), target texts (TT) and reference texts (see Sect. 3.3 for the categorisation of features as contributing to different translationese effects).

The two interpretations of the nature of translations given by Toury and by Baker are complementary and can be seen to represent two major types of translationese. To avoid unnecessary associations with the foreign language acquisition terminology, we would use Elke Teich's term *shining-through* to refer to the cases where the cross-linguistically diverging frequencies of the features are adapted in translations to the SL values, giving rise to significant distinctions between translations and non-translations (Teich 2003). This is the 'interference' type of translationese, which is considered the major factor in shaping the properties of translations (see evidence in Evert and Neumann 2017; Volansky et al. 2015, for example). The features of translations that significantly deviate from both SL and TL, where there are no cross-linguistic differences between non-translations (English source texts and originally-authored Russian texts in our setup), should be considered cases of true *language-pair-independent translationese* in line with Baker's ideas. Some features that spot language contrast can be fully adapted to the TL norm (*adaptation*) or even exaggerate the TL properties (*over-normalisation* or russification in our setup).

Methodological paradigms in translationese studies (features, data and analytical approaches). Over the last few decades, translationese studies as an area of research within translation studies has seen significant developments in the research methods. The earlier investigations were often based on manual extraction of a few features from limited corpus data (sometimes lacking the parallel component) and relied on univariate statistic analysis (Becher 2011; Castagnoli et al. 2011; Nakamura 2007; Puurtinen 2003; Santos 1995). The more recent projects are computationally intensive and involve massive parallel and comparable corpus resources in several language pairs and complex research designs with extensive and elaborate feature sets and methods (see, for example, Dipper, Seiss, and Zinsmeister (2012) who describe the typical corpus resources setup in translationese studies and Evert and Neumann (2017) for the multivariate analysis and feature engineering methodology).

A *machine learning (ML) turn* in the translationese research began with the ground-breaking work by Baroni and Bernardini (2006) who convincingly demonstrated that translations of geopolitical texts into Italian are inherently different from the comparable non-translations by employing a Support Vector Machines (SVM) algorithm to classify them. They experimented with various types of n-grams to represent texts and discovered that bigrams performed best. An important message from their experiments was that a ML algorithm was able to reliably pick the difference between translations and non-translations even when the human subjects (professional translators) were unable to do so as effectively. It brought about a new strand of research known as translationese detection. ML algorithms were used to test the hypothesis about various translationese properties. A good example of this methodology in action is Koppel and Ordan (2011), who reported a series of ML experiments on the Europarl corpus and confirmed that source language plays a crucial role in the make-up of a translated text. They used frequencies of 300 function words as features (which excludes any cultural or topic differences between the corpora). Probably, the most impressive results were

reported by Popescu (2011) who reported 99.53% cross-validation accuracy in the task of detecting translations on character string features for an SVM classifier trained on literary translations from French and German into English. However, when they tested a model trained on out-of-French translations on out-of-German translations they received the results at the chance level—an indication that character n-grams capture uninteresting SL-related cues such as proper names. Filtering out those items led to the realistically moderate results of 77.08% in the experiment where they trained on translations from French and books by British authors for reference and testing on translations from German and American fiction for non-translated reference.

In Ilisei et al (2010), a supervised learning approach was employed to identify the most informative features that characterised translations compared to non-translated texts. The learning system was trained on two domains, medical and technical. The novelty of their approach consisted of its language-independent data representation. On the categorisation task, the algorithms achieved an accuracy of 87.16% on a test set and reached up to 97.62% for separate test datasets from the technical domain. The removal of the features, linked by the authors to simplification, from the machine learning process led to decreased accuracy of the classifiers. Therefore, the retrieved results were interpreted as an argument for the existence of the simplification universal.

The book by Gloria Corpas presents the results of several NLP experiments to study translation universals and translationese features. Corpas focuses on three universals: simplification, convergence and transfer (shining-through). Vectors of lexical and syntactic features are used to test various corpora of English and Spanish: (a) a large corpus of Peninsular Spanish (reference corpus of 50 million words), and various comparable corpora: (a) corpus of translation of medical texts by professionals and semi-professionals (from English into Spanish); (b) corpus of non-translated medical texts in Spanish; c) corpus of non-translated medical texts in English, (d) corpus of translation of technical texts by professionals (from English into Spanish); and (d) corpus of non-translated technical texts in Spanish. The main findings support (1) the inexistence of simplification of translated text into Spanish (for most features) (non-translated Spanish texts are even more simple). (2) Convergence (translated texts are more homogeneous among themselves) can be observed only for syntactic features. (3) Transfer can only be observed partially: there is some positive transfer (translated texts show more lexical cognates), but no negative transfer (translated texts show more zero pronouns). Syntactic interference (shining-through) is observed for all translated texts (Corpas Pastor 2008).

After the initial sweeping success of ML approaches to detecting translations on surface and linguistically uninterpretable features, there appeared a research strand that aimed to combine the ML computational power with the corpus-linguistic interest in translationese properties. These efforts can be exemplified by Volansky, Ordan, and Wintner (2015) research, which tested the usefulness of a dozen of linguistically informed features, theoretically attributed to the main translation tendencies (simplification, interference, normalisation and explicitation). In effect, they used ML methodology to perform univariate analysis (they compare the

accuracy of a binary translationese classification on each feature) to reveal the features prominence in the identification of translations. Their findings make a strong argument for interference as the major tendency in translation and, concurrently, for language-pair-related nature of translationese in general. The authors also make rigorous claims about the importance of a parallel data, content-independent features and genre-related nature of translationese trends.

The use of automatic text classification as a validation methodology combined with unsupervised and mildly supervised machine learning techniques (namely, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)) was promoted in Evert and Neumann (2017) for revealing the latent distinctions between text types (languages, registers, translations-non-translations) and exploring the sets of features that load on the respective discriminants. Unlike the previous research, the authors advocated the use of the *multivariate techniques* claiming that translationese is a systematic property of a text, not dissimilar to register specificity, and can hardly be conveyed by a single feature, but rather a combination of them (cf. multidimensional approach to register studies introduced by Biber (1988) and similar approach to translation in (Prieels et al. 2015)). An important methodological claim that the authors make is about the resources necessary for translationese studies. They assert that 'it is methodologically impossible to determine differences between translated and non-translated texts without comparing the realisation of a feature in the matching source text' (Evert and Neumann 2017: 49). It is interesting to note that despite their study is based on a balanced corpus involving five registers, the register variation was treated as a confounding factor that shapes translationese; any register-related interpretations were left for future work.

Evidence for language pair specificity of translationese. While developing effective language-independent applications to detect translations can be an interesting engineering task, there is ample evidence that translationese features and effects are indeed language pair and translation direction specific. In fact, the symmetric additions and omissions of items in both translation directions between two languages (demonstrated by Becher (2011), for example) are indicative of the impact of the contrastive properties of the language pair on the translators' choices. Reduced accuracy of the translationese classification, when a model trained on translations from SL1 is tested on translations from SL2, supports the same conclusion (Koppel and Ordan 2011; Popescu 2011). It is common to interpret the linguistic make-up of translations as a complex interplay of the two major forces: the SL shining-through pull and the TL normalisation pull (see, for example, Hansen-Schirra (2011)).

To sum up, the previous translationese research has established that translations are systematically and inherently different from the originally-authored texts due to the specificity of the underlying communicative situation and cognitive processes. It has been shown that the property of 'being a translation' is largely determined by the SL and the register conventions. The intuitive association between some frequency features and translationese universals proved difficult to be confirmed by empirical evidence due to the lack of objective link between the trend and its operationalisation. However, bottom-up exploratory approaches based on ML

methods enable to reveal translationese indicators and the unique ways in which they coalesce into patterns in each register of a given translation direction.

In general, the relevance of translationese studies is supported by the renewed interest to the impact the human translated training data exerts on the quality of machine translation (Aharoni et al. 2014; Goutte et al. 2009; Graham et al. 2019; Popovic 2020; Stymne 2017; Zhang and Toral 2019). One of the earlier investigations into this issue by Lembersky, Ordan, and Wintner (2012) demonstrated that the BLEU score can be improved if the language models are trained on the translated texts and not the texts originally written in the TL.

The current project is based on balanced data for four registers, each represented by a combination of (1) a document- and sentence-aligned parallel corpus of professional published translations for English-to-Russian language pair and (2) a comparable corpus of non-translations in the target language. These components are necessary to reliably capture and describe various translationese effects by comparing feature frequencies across three text types in each register: sources, targets and reference texts. Methodologically, we combine multivariate analysis in supervised and unsupervised ML settings and univariate statistical analysis to reveal prominent translationese indicators and describe trends observed within and across the registers. Our features include content-independent morphosyntactic features that allow to abstract from topic and domain information as well as indirect lexical indicators retrieved from language models learnt on separate and much bigger register-comparable resources. Importantly, all features are shared by the two languages involved to enable placing all texts into the same multidimensional feature space.

2.2 *Translationese and Register*

This research explores the translation properties that are observed in various registers. It is difficult to deny that language is not homogeneous. Language is a combination of subsystems that are employed in specific communicative conditions. One important dimension of language variation, distinct from domain sublanguages, territorial or social dialects, has to do with the dominant communicative function and the generalised type of the situation in which the textual activity takes place. This type of variation is referred to as registers or genres depending on which aspects of the communicative event are focused. David Lee, the author of one of the text categorisation schemes in the British National Corpus (BNC), prefers to think about these competing terms as 'two different points of view covering the same ground' (Lee 2001: 46). The term *register* signals that language material is approached from the viewpoint of its internal properties (such as frequencies of linguistic items), which form specific patterns of use predetermined by the communicative conditions ('the context of the situation') in which they occur. The major situational factors are typically described following Halliday's categorisation into field, tenor and mode. Genres are understood as text categories more focused

on the text-external and functional parameters; they are text schemata licensed by the culture and superimposed on the register. According to James Martin, 'no culture combines field, mode and tenor variables freely' (1992: 562). This approach is in line with Michael Halliday's interpretation of register (see Register Variation chapter in Halliday and Hasan 1989) and is adapted in a number of corpus and computational linguistics projects, especially based on the BNC (see Lijffijt et al. 2016; Neumann 2013; Santini et al. 2010; Sharoff 2018).

In translationese studies, it seems more typical to refer to the analysed text categories as registers (see Diwersy et al. 2014; Kruger and Rooy 2012; Lapshinova-Koltunski 2017 among other works). However, Delaere (2015) consistently prefers the term 'genre' to refer to the text categories of similar names and granularity, because in her research these categories are explicitly annotated using such non-linguistic characteristics (addressor, addressee, channel and communicative purpose), following the methodology in Biber and Conrad (2009).

In the current research, we follow this interpretation of the contextual language variation and refer to the four text categories under comparison (general domain mass-media texts, popular-scientific texts, fiction, political-economic news commentary) as registers.

Register is widely acknowledged as one of the major factors that influences the properties of translations, along with the source language.[2] This is not surprising precisely because of the strong SL pull in translations, given that 'parallel registers are indeed more similar cross-linguistically than are disparate registers within a single language' (Biber 1995: 279). In a lot of earlier research, this is corroborated as a by-product of a different research focus and/or as a result of observations from manual analysis of some restricted corpus data. For example, a relatively small-scale study based on half-a-million word corpus by Puurtinen (2003) indicated that genre could be an important factor guiding translation choices. The authors concluded that 'subgenres of children's literature … should be investigated separately' (Puurtinen 2003: 403).

Xiao, He, and Ming (2010) report the construction of a register-balanced corpus of translational Chinese and original Chinese texts after the FLOB sampling frame. In their univariate analysis of several known translationese indicators, they show that the features tested, including lexical density (STTR), mean sentence length, conjunctions and passives frequencies, display 'genre subtleties' in translation.

Our research can be compared to Kruger and Rooy (2012), who see the investigation of the relationship between register and the features of translated language as one of their main research goals. They performed univariate analysis for seven features, which represented three translationese universals, to see how the universals would play out within and across their six registers. In their research design, explicitation, normalisation and simplification were operationalised with the (1) frequencies of full forms (as opposed to contractions), that-complementisers,

[2]Earlier studies that suggest that translationese is dependent on register are Steiner (1998), Reiss (1989) and Teich (2003), among others.

linking adverbials; (2) frequencies of coinages, loanwords and common lexical bundles; and (3) values for lexical diversity and mean word length, respectively. Their results provided limited evidence for universal character of translationese, rather each register demonstrated its own pattern of analysed features. In a later research using the same features, the levelling-out of registers, conceptualised as the assumed reduced register variability in favour of a neutral middle register, was not supported either (Redelinghuys 2016).

In recognition of the importance of register in translationese studies, researchers pay special attention to the selection and annotation of the reference corpus of non-translations: Castagnoli (2009) decided to build a new corpus from scratch, Delaere (2015) re-annotated an existing resource, Kunilovskaya and Lapshinova-Koltunski (2020) used a special corpus sampling strategy to extract functionally comparable subsets from larger corpus resources.

The large-scale studies of translated registers that allow reliable application of statistical methods or ML techniques are comparatively rare. There is a case study in Diwersy, Evert, and Neumann (2014), based on a reasonably large register-balanced bidirectional English and German corpus, but its contributions were more of the methodological nature: they reported few findings that characterised individual registers in translation, if any.

Delaere (2015) used the frequencies of linguistic items associated with the general properties of texts such as formal/neutral language and native/borrowed words to profile originally-authored and translated texts and test whether the translators tend to conform to the observed TL norm. Her findings for five genres in several language directions between Dutch, English and French generally confirmed the normalisation trend in translations and the impact of the genre and SL factors, but there was no consistency in the results. The authors attributed this inconsistency to incomplete metadata in the corpus and some unaccounted factors that might govern translators' choices. The sparsity of the indicators and domain disparities could also be confounding factors, given the lexical nature of the operationalisations implemented and the relatively small size of each subcorpus used in the study.

Unlike the previous study, which relied on predefined operationalisations of some properties of translated texts like levels of formality, Lapshinova-Koltunski (2017) employed hierarchical cluster analysis, an unsupervised ML method, and represented English-to-German translations and German non-translations in seven registers as feature vectors using eight lexico-grammatical patterns that were inspired by register studies to see how much the properties of translations were influenced by two factors—the register and the method of translation. Their features are context-independent and characterise texts through ratios of, for example, nominal vs. verbal parts-of-speech or through cumulative frequency values for items expressing modality or evaluation among others. The results of the study showed that the functional text type dimension dominated as a factor for some registers but not others. This research, as well as an earlier research on the same data using SVM classification (Vela and Lapshinova-Koltunski 2015), had its focus on the comparison of human and machine translation across a range of registers.

They found that the two translation varieties were more similar between themselves than any of them were similar to the register-comparable non-translations. In a later work on the same data, they used part-of-speech (PoS) trigrams in a number of binary text classification experiments to reveal and interpret features distinguishing translated registers. They confirmed their earlier finding that 'the genre dimensions in translation variation is much stronger than that of translation method' (Lapshinova-Koltunski and Zampieri 2018: 107). These three studies indicate that human and machine translations are more similar between themselves than any two translated genres, regardless the feature set used and ML approach chosen.

3 Methodology

In translationese research, the results are largely dependent on the features used to represent the texts, including their selection and extraction. Features are usually frequencies or ratios of linguistic items and phenomena, used to operationalise various hypothesised translationese trends or to capture and measure translationese effects in the bottom-up approach.

Another important factor is the type, quality and size of the corpus resources used to produce data tables. As it is shown above, both parallel and comparable components are required to be able to interpret quantitative differences between translation and non-translations.

There can be various ways of looking at the data methodologically, ranging from manual in-depth analysis of a few contrastive linguistic phenomena and/or statistical significance testing to ML experiments, usually cast as text classification problems or various types of factor analysis and computational linguistics methods. While the previous research has reported some tried and tested approaches, they leave a lot of room for development and exploration, especially if new research questions are posed.

Unlike much of the related work, where register effects on translationese properties are used as a backdrop for another primary research questions, the current research employs ML techniques to compare the type and strength of various translationese effects in several registers as well as to reveal the translationese indicators that might cut across all registers. This section has the description of these three major components of our research design: features, data and methods.

3.1 Feature Sets

Similarly to Volansky, Ordan, and Wintner (2015), our features are not selected to get the highest accuracy for the binary classification of originally-authored texts and translations (translationese classification). We seek to investigate the variation in translations along the register dimension in a linguistically interpretable way.

In the literature, the types of features used to capture translationese in the ML setting vary depending on the specific task. Translationese detection and SL identification tasks almost exclusively rely on character, word, lemma, PoS or mixed n-grams of various order[3] and most frequent lemmas (including function words) or PoS.[4] A bold exception is the projects that aim at sentence-level detection of translation direction (Eetemadi and Toutanova 2015; Sominsky and Wintner 2019). They leverage the aligned PoS information from source and target sides of the parallel corpora to achieve the state-of-the-art results. Sominsky and Wintner (2019) reported further improvements of up to 6% accuracy (at the expense of interpretability) for four out of six tested language pairs on distributional 50-dimension pre-trained GloVe word embeddings used to represent words and fed to a neural network of one bidirectional Long Short-Term Memory (BiLSTM) layer.

The more linguistically orientated research, which aims to know more about the linguistic specificity of translations, considers the feature selection the most challenging and creative part of the task. On top of the well-known and most-tested translationese indicators (such as type-to-token ratio, content-to-function words ratio, frequency of connectives/conjunctions and pronouns, ratio of contracted to full forms, average sentence length, mean word rank), the authors suggest more elaborately engineered features. For example, Arase and Zhou (2013) used the frequency of discontinuous structures to capture 'phrase salad' in MT. Redelinghuys (2016) calculated readability scores, while Volansky, Ordan, and Wintner (2015) operationalised the normalisation hypothesis with average point-wise mutual information (PMI, one of the association measures used to detect collocations) of all bigrams and ratio of repeated content words along with other features. Lapshinova-Koltunski (2017) suggested a feature set, which included features like frequency of evaluative patterns and degree of nominalisation (ratio of nominal and verbal PoS). Some experimenting was done with the frequency features based on parsed data: Ilisei et al. (2010) calculated ratio of simple sentences and parse tree depth and Kunilovskaya and Kutuzov (2018) extracted and counted syntactic relations tags from UD annotations of their corpora.

In our research the feature selection and engineering process was informed (1) by the findings in the translation and translationese studies, including the practical observations made in English-to-Russian translation textbooks, but never tested empirically and (2) by the practices in the register studies and variational linguistics on the assumption that translations could be viewed as a specific sublanguage, a third code (Duff 1981; Frawley 1984), based on the specificity of distribution of the linguistic features. This is supposed to enable measuring the cross-linguistic distance between the registers as well as between translations and non-translations. This approach effectively means that our feature set is language

[3]See, for instance, Baroni (2006), Kurokawa (2009), Arase (2013), Eetemadi (2015) and Rabinovich (2016).

[4]Some relevant studies are Popescu (2011), Koppel (2011) and Nisioi (2013).

pair specific and would require adaptation to be extended to other language pairs (see such adaptation in Kunilovskaya and Lapshinova-Koltunski 2020). Besides, our research design required that the features (3) should be shared by the languages involved in the experiment. We also focused on (4) content-independent features to reduce the noise from the topic and domain divergence between the parallel and the reference corpora, which excluded the common bag-of-words models from our options. Finally, we avoided (5) less interpretable features and (6) features that defy reliable extraction based on our experience.

Unlike much of the previous research into translationese, overviewed in Sect. 2.1, we do not assign features to the known translationese trends in the top-down manner, but empirically establish their role in producing various translationese effects. The experimental setup in this study can handle irrelevant or collinear features, and we use a reasonably high number of potential translationese indicators to be able to distil the most useful ones through feature selection.

Our feature set is composed of two parts. First, it includes 45 morphosyntactic features that were introduced in Kunilovskaya and Lapshinova-Koltunski (2019) to capture human translation quality. We provide a brief overview of these features below. For the full description of each individual feature, refer to Appendix. The feature codes used in this chapter and the extraction details are given in the Appendix alphabetically. Second, it comprises 11 abstract lexical features to reflect the specificity of the lexical choice in translations.

The morphosyntactic features are extracted from the annotation performed within the Universal Dependencies framework (Straka and Straková 2017), using models pre-trained on 2.5 versions of the EWT and SynTagRus treebanks for English and Russian, respectively.

More than a third of these features (17) are the frequencies of the default UD morphosyntactic tags (such as *ccomp*: clausal complements or *sconj*: subordinating conjunctions) and their combinations (such as *numcls*: number of clauses per sentence counted as the number of relations tagged *csubj, acl:relcl, advcl, acl, xcomp* in one sentence); when extracting PoS tags for various types of pronouns and other closed word classes, we used lists to filter out noise. The other third of the features (16) involved custom rules and extraction patterns, detailed in Appendix. These include *lexical type-to-token ratio, modal predicates, passives, mean dependency distance* (*mdd*, which represents 'comprehension difficulty' defined as 'the distance between words and their parents, measured in terms of intervening words' (Jing and Liu 2015). In developing these features we took into consideration the description in (Evert and Neumann 2017; Nini 2015) for English and in (Katinskaya and Sharoff 2015) for Russian. Further on, the cumulative frequencies for the four semantic types of *connectives, epistemic markers* and *adverbial quantifiers* are extracted using predefined lists compiled from the literature (see more details on the items selection, academic sources, extraction and disambiguation in Appendix).

Generally, our UD-based indicators include morphological forms (e.g. non-finite forms of verbs), syntactic relations (e.g. clausal complements), syntactic functions (e.g. modal predicates), word classes (e.g. pronouns, discourse markers). The extraction quality of these features largely depends on the quality of the UD

annotation: for v2.5 mean accuracy on raw text is reported at 93.3/97.8 for universal PoS, 94.2/93.5 for morphological features and 77.0/85.0 for labelled dependency attachment for English/Russian, respectively.[5]

For this project we implemented 11 additional features to approach translationese at the lexical level as well. It is obvious that we cannot rely on frequencies of individual character or word n-grams in our cross-lingual setting. Besides, it is a known fact that sparse vectors of string features do not generalise well across domains (Eetemadi and Toutanova 2015). Instead, we used language model (LM) perplexities and calculated ratios of n-grams from top and bottom frequency quartiles, using the KenLM toolkit (Heafield 2011) and Quest ++ utilities (Specia et al. 2015). These features are used for the analysis of translationese in the research projects, which target translation quality (see Karakanta and Teich 2019 and Quest ++ feature set). We hypothesise that translated texts might have a diverging lexical composition in terms of ratios of n-grams from high- and low-frequency bands and sentence perplexity scores due to unseen sequences induced by the translation process. Our text-level lexical features include:

- mean target sentence perplexity score from the 3-g language models trained on large register-comparable corpora (see 3.2.2 for details);
- standard deviation value for the above sentence perplexities to account for possibly uneven lexical complexity of sentences in the translated texts;
- ratio of uni-, bi-, trigram that were not seen in the n-gram lists from the reference corpora;
- ratio of n-grams from the 1st frequency quartile (low-frequency items)
- ratio of n-grams from the 4th frequency quartile (high-frequency items)

To produce these features, we collected separate language resources for each register making sure they do not intersect with the smaller reference corpora included in our experimental data to exclude unfair bias for these features. Before learning LMs and generating n-gram lists, all corpora had been lemmatised and PoS-tagged with UDPipe (Straka and Straková 2017) to get lempos representation (e.g. as_SCONJ i_PRON look_VERB up_ADP ._PUNCT). This is required because Russian is a morphologically rich language; English is pre-processed for higher consistency and comparability.

As a result of feature extraction, each text in our data was represented as a vector, where individual components corresponded to the value of each feature for this text. The dataset, used in the experiments, can be thought of as a table, which has texts in rows and features in columns. Note that prior to the experiments, the values of each feature were standardised to get the distribution with a mean value 0 and standard deviation of 1. This helps to ensure that all features have the variance of the same order, and each feature makes the same contribution to the differences observed, regardless of large discrepancies in real values between some indicators.

[5] http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_20_models.

3.2 Research Corpora

This research relies on several parallel and comparable corpora to explore the linguistic properties of texts translated from English into Russian by professional translators across a variety of registers. We distinguish between the corpora used to conduct experiments (data) and the corpora used to learn language models and produce n-gram frequency lists (linguistic resources).

All corpora were put through the same pre-processing pipeline (spelling unification, text size normalisation, deduplication, noise filtering), annotated with UDPipe and converted to PoS-tagged lemmas (lempos format).

3.2.1 Data

The selection of registers for this project was limited by the availability of the English-Russian parallel and comparable corpora that would store texts of reasonable size and structure. We considered a wide variety of the available parallel corpora, including web corpora (Yandex 1 M-token parallel corpus, Parallel Corpora for European Languages), United Nations corpus, corpora of subtitles and Wiki Titles, TedTalks corpora and mozilla transvision corpus of technical translations. But the units of storage in these corpora were often limited to one sentence or would include a lot of non-textual information and tables. TedTalks transcripts and subtitles have specific translation processes behind them that can unfairly influence the frequencies of our features. It is also more difficult to make assumptions about the translation quality for these corpora and compile non-translated comparable corpora for them.

We focused on the four registers: general domain mass-media texts, popular-scientific texts, fiction and the news commentary texts in the political and economic domain. All translations included in the experiments are published. We only selected the corpora that store texts with respect to their natural text boundaries, which allows the collection of text-level statistics. The parallel subcorpora are document-level and sentence-aligned. The global sources of data in this project can be described as follows.

1. *Mass-media* parallel corpora include data from the three major sources: a quarter comes from the parallel component of the *Russian National Corpus (RNC)*[6] and the rest of the data were manually collected or crawled from *InoSMI.ru* and *BBC.com/russian* (2018–2020).
2. *Popular scientific* parallel corpus is self-compiled from a dozen of full-length English books on a range of subjects including biology, physics, sociology, history, anthropology, robotics, medicine, and their published translations into Russian from 1999 to 2016 period. This corpus is now included into the RNC

[6] https://ruscorpora.ru/.

Table 1 The macro-corpus used for research purposes (k=thousand, m=million)

	Type of data	Words	Sentences	Documents
general media	parallel	731 k	31 k	525
	reference	625 k	33 k	448
popular science	parallel	1 m	42 k	112
	reference	1 m	46 k	101
fiction	parallel	11 m	564 k	149
	reference	12 m	706 k	200
commentary	parallel	301 k	12 k	347
	reference	276 k	13 k	334

parallel resources. While the number of observations is small, the selected unit of storage is a chapter or a part of the book.

3. The parallel data for *fiction* is entirely from the RNC parallel component. It includes 149 source texts of various length and literary genres, but mostly novels representing over a hundred of authors from Dickens to Rowling.
4. Parallel *political and economic articles (commentary)* are extracted from the *WMT News Commentary* corpus (v.15),[7] which contains political and economic commentary crawled from *Project Syndicate* website.

The originally-authored Russian texts to be used as the reference for the former three registers were randomly sampled from the respective register subcorpus of the main 500-million RNC and for the last category—from the 300-million contemporary Russian newspaper corpus, included in the RNC monolingual resources.

Table 1 has the description of the pre-processed and annotated parts of our register-balanced corpus including the parallel and comparable monolingual components. For the parallel data we report the size on the SL side only.

In total we have 3349 documents in two languages, labelled for four registers and three types (sources, targets, reference).

3.2.2 Linguistic Resources

The resources for LM training in all registers, except the English news commentary, come from the British National Corpus (BNC) and the Russian National Corpus (RNC). We relied on the available metadata to ensure maximum comparability with the parallel data in terms of intended audience, text production time and communicative function. The English political and economic commentary reference texts are collected from the WMT News Commentary corpus outside the English-Russian parallel data. Note that these resources exclude the random

[7] http://www.casmacat.eu/corpus/news-commentary.html.

Table 2 Corpora used to train language models and generate n-gram lists

	Language	Words	Sentences	Documents
general media	en	3.9 m	177 k	100
	ru	129 m	6.9 m	226 k
popular science	en	17.7 m	682 k	528
	ru	1.9 m	93 k	378
fiction	en	18.6 m	1.2 m	431
	ru	37.6 m	2.6 m	580
commentary	en	5.9 m	237 k	8.7 k
	ru	5.7 m	252 k	9.5 k

samples used as reference data and described in Table 1. The general shape of the resources after pre-processing and annotations can be found in Table 2.

We will indicate that the mass-media items in the BNC do not observe true document boundaries but are in fact text chunks of varying length. However, it is irrelevant for the purposes of building LMs and n-gram lists.

3.3 Methods

Our methodology combines the data representation and visualisation approaches which were shown to be effective for the study of translations in Evert and Neumann (2017) and the idea that in revealing or measuring translationese effects, the distance between the source and target languages (or, in our case, registers) has to be taken into account. We develop the general approach tested in Kunilovskaya and Lapshinova-Koltunski (2020) on one register for two language pairs.

To represent texts in our data we generate feature vectors, where each component has the value for a particular linguistic parameter. With the exception of the LM perplexity scores, these values are the frequencies or ratios of a targeted linguistic phenomenon, captured through a set of PoS tags or a syntactic pattern. For features based on the search lists, the values are cumulative frequencies of all items on the respective list. For n-gram counts, we used an empirically established frequency threshold of 10, which means that we ignored the n-grams with a frequency lower than 10. This measure helps to avoid zero values for bigram and trigram ratios. Given that our features are the same for all text categories and text types, this representation effectively puts them in a shared feature space. The extraction details are given in Sect. 3.1 and in Appendix.

We resort to PCA, an unsupervised ML technique, for dimensionality reduction to present our observations in scatter plots and visually estimate whether our features reflect the ontological text categories and types. The visual impressions are verified by the results of text classification. In all experiments we rely on the linear SVM algorithm, set to the default scikit-learn parameters (C = 1.0, degree = 3, gamma = auto). The algorithm is fed with the feature vectors that have been centred

around the mean and scaled to unit variance and is run in the 'balanced' mode to offset the unequal number of observations in the training classes. We report the results in the tenfold cross-validation setting to reduce the possible biases of any single held-out test set.

In accord with our research questions, given in Sect. 1, the text classifications are designed to capture the following general properties and phenomena:

- translational status: a binary classification for each register;
- register variation: a 4-label classification for non-translations in each language;
- standardisation effect: a 4-label classification for translated texts only.

To determine the position of each translated register with regard to the sources and TL non-translations, we average the real-valued vectors across each of the three text types and calculate the *Euclidean distances* (a square root from the sum of squared differences between the corresponding dimensions of the two vectors) between them. We rely on the Euclidean distance (as opposed to cosine similarity, for example) because in this experiment we use unscaled vectors and the magnitude of the values in each dimension matters. The differences between the three measurements, which can be pictured as triangles, demonstrate the relative proximity (similarity) of the translated texts to the originally-authored registers in the two languages. The idea to measure linguistic (morphosyntactic) distances between languages for the purposes of translationese studies is not new. To this end, Nikolaev et al. (2020) computed the cross-linguistic congruence index as the proportion of matching universal PoS tags and dependency labels for all manually aligned content words in a parallel corpus. They acknowledged that there was no established procedure to achieve it.

The explanatory analysis of the linguistic specificity of translations in each register is based on the best translationese indicators, i.e. the top N features that can be used by the ML learning algorithm to differentiate the classes with the minimum loss in the classifier performance. Our experimental results indicated that the best performance for the top 10 and top 20 features was returned by the *Recursive Feature Elimination (RFE)* feature selection algorithm, which internally used *Support Vector Regressor (SVR)* with the default scikit-learn settings. The same approach was used to reveal register contrast indicators that were necessary to demonstrate the amount of intersection between the translationese and cross-lingual contrast features.

Finally, we perform a succession of the univariate analyses to establish which features contribute to various translationese effects that we distinguish in this study following a procedure described below. In all experiments we used the *two-tailed T-test for samples with unequal variance* and quantified the effect size of the differences with *Cohen's d*. First, we identify the features that have significant differences between translations and non-translations (tgt, ref): these are translationese indicators. Then, we establish whether there are differences between the two cross-linguistic registers (src, ref) with respect to a given feature (the language gap). Finally, we compare the average frequency for the feature in translations with those

in the source and target languages to determine how it relates to these values (greater or smaller).

Combinations of these tests outcomes yield the feature sets for the following translationese effects:

1. shining-through effect: translationese features in the language gap, i.e. we observe significant differences between translations and non-translations and between English and Russian non-translations; and the frequencies of features from translations are smaller than in English but significantly greater than in non-translated Russian (src > tgt > ref) or greater than in English but smaller than in Russian (src < tgt < ref);
2. anglicisation: translationese features demonstrating frequencies outside the English extent of the significant language gap;
3. SL/TL-independent translationese: translationese features with significant differences from both languages and no language gap;
4. over-normalisation: translationese features demonstrating frequencies outside the Russian extent of the significant language gap;
5. adaptation: features that have significant differences for the two languages, but not translationese features, i.e. their frequencies are adapted to the TL norm.

This procedure is also supposed to reveal features that are useless for our purposes: the feature that has the same frequencies in translations and non-translations, and also do not distinguish the languages.

4 Results

In this section, we first report the results of the two classification experiments that test the ability of our feature sets (1) to distinguish translations and non-translations in each register, (2) to capture the register variation in the originally-authored texts in each language. We also look at the performance of the register classification on the translated registers to check whether the register distinctions are diluted by the translation process. If the translated registers are more difficult to classify, we can confirm the levelling-out hypothesis. The second paragraph demonstrates how the translated registers are positioned against comparable non-translations in both languages (src, ref) based on the Euclidean distances in our setup. We complement the spacial representation of translated and non-translated registers with histograms for values on the strongest PCA dimension, which appears to mostly capture register variation in our data. Finally, we describe the subsets of features that are revealed through feature selection and comparative frequency analysis and represent several translationese effects. Feature analysis is performed to explain the observed specificity of each translated register with regard to their sources and reference non-translations.

4.1 Translationese and Register Distinctions

For a preliminary investigation of the data, given our features, we visualised the distinctions between all text types on the full feature set and on its morphosyntactic and lexical parts. For example, Fig. 1 has a scatter plot, where each document is represented by the values on the first two PCA dimensions, i.e. the result of the dimensionality reduction of the 45-dimensional morphosyntactic vector. Unlike lexical features (not shown for the consideration of space), the morphosyntactic features manage good separation of the registers and the two languages. It seems that the register variation is found on Dimension 1, which explains the most variance in the data, while Dimension 2 (shown on the vertical axis in Fig. 1) captures the language contrast. The lexical features are not able to achieve this representation of data on the most prominent known properties of the texts: they squeeze all variance into the first dimension. It means that in terms of ratios of high-frequency and low-frequency n-grams the similarity between registers from different languages is stronger than the differences between languages. This observation is confirmed by the language contrast classification (English vs. Russian original texts) results: for morphosyntax 100% accuracy can be achieved

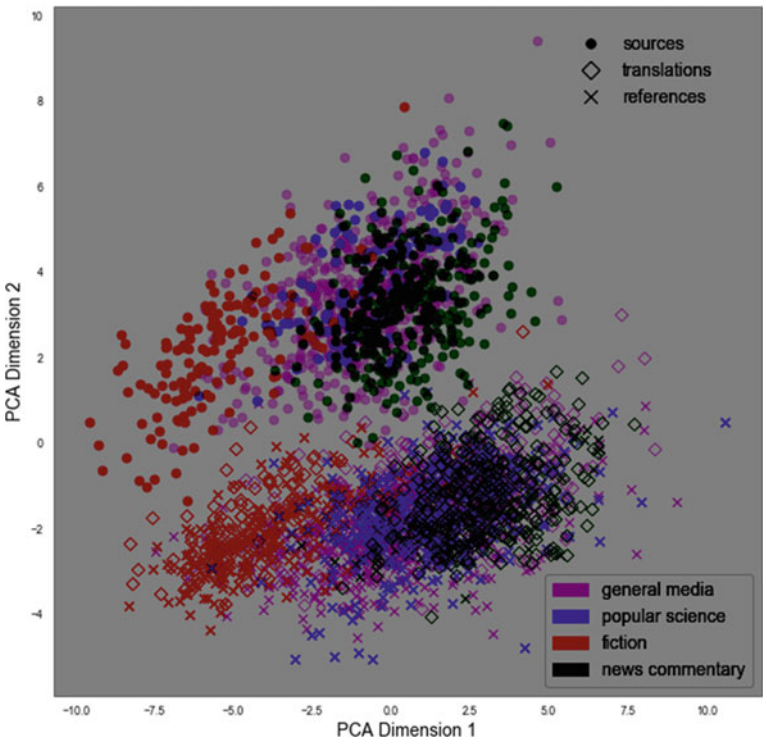

Fig. 1 Values on the first two PCA dimensions derived from the morphosyntactic features

on just 3 features (*aux, aux:pass, parataxis*), while the 11 lexical features returned only 85%.

The concatenation of the two feature sets captures the register distinctions on Dimension 1 and language distinctions on Dimension 2 more clearly (see Fig. 3).

However, the distinctions of translations and non-translations, required by the first step in our methodology, are clouded. To bring them to the fore for closer exploration, we tried to cast the full feature vectors of size 56 for translations and non-translated Russian texts to a bidimensional space by PCA and produced a scatter plot of the resulting data. The independent subplots in Fig. 2 position the texts in each register according to the values received on the first two principle components.

It can be seen that translations are shifted away from the non-translations, especially in general mass media and news commentary. It means that our features do register some divergence of translated Russian from the expected TL norm in these registers represented by non-translations. Admittedly, the visual impressions are more subtle in the other two registers. Note that PCA is unsupervised: it is unaware of any text types that are colour-coded in the plots. Besides, PCA reduces the 56 dimensions to just two, necessary to plot the data, which inevitably leads to the loss of information and distortions. That is why we verify the visual impressions

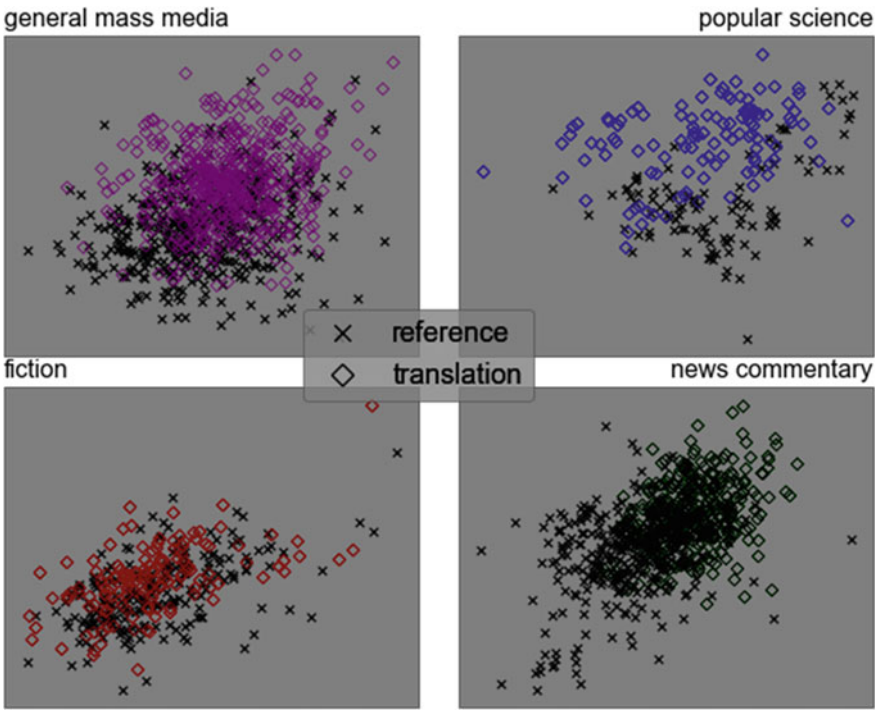

Fig. 2 Differences between translations and non-translations by register

Table 3 SVM performance on the translationese classification in each register

	N texts	56 features		45 morphosyntax		11 lexis	
general media	973	87%	0.872	**87%**	**0.869**	75%	0.750
popular science	213	98%	0.977	**98%**	**0.981**	76%	0.754
fiction	349	95%	0.953	94%	0.936	77%	0.759
commentary	681	95%	0.947	93%	0.934	**88%**	**0.879**

with a series of binary translationese classifications using SVM. The classification results confirm that PCA visualisations can be, indeed, misleading, because the registers with seemingly different visual distinctions (fiction and news commentary) achieve the same high classification accuracy, while the accuracy for general mass media is lower, in contrast with what is observed in Fig. 2.

The cross-validation results are presented in Table 3, which shows SVM performance on the translationese classification, taking into account accuracies and macro F1 scores. On the full feature set in three registers, SVM achieves the accuracy of over 95%, while for mass-media texts it is 87%, which is still reasonable high. We have fairly balanced classes in all registers, so the chance level never exceeds 50%.

The classification experiments on morphosyntactic and lexical feature sets separately indicate that the result in the 56 features column (see Table 3) is mostly produced by the morphosyntactic features. If lexical features are eliminated the classifier performance does not degrade much in any registers: the loss amounts to 1% and 2% in accuracy for fiction and commentary at most. However, switching to just lexical features results in the drops in performance ranging from minimum 7% (*news commentary*) to maximum 17% (*popular science*). It means that for the translationese classification (1) *news commentary* relies on the lexical features most, i.e. they demonstrate the highest divergence from non-translations; (2) for *popular science* structure is most important, i.e. translations differ from non-translations in morphosyntax; (3) in *general media* both feature sets perform the worst, possibly because of the higher variation in the respective subcorpora observed in Fig. 3.

Secondly, we are interested in finding out whether our features model the register diversity in both non-translated languages well. In Fig. 3 we plotted the originally-authored texts in the two languages, represented by their values on the first two PCA dimensions generated by the PCA transform of the full feature vector of size 56. Most variance is explained by Dimension 1, which captures register variation. Texts from different registers seem to occupy specific areas along the horizontal axis, especially in Russian. The second dimension has the clear separation of the two languages. The plot in Fig. 3 also indicates that some eponymous registers are closer together across languages than others. For example, *fiction* and *news commentary* seem to be more similar along the vertical 'language contrast' dimension than *general mass-media texts* and *popular science*.

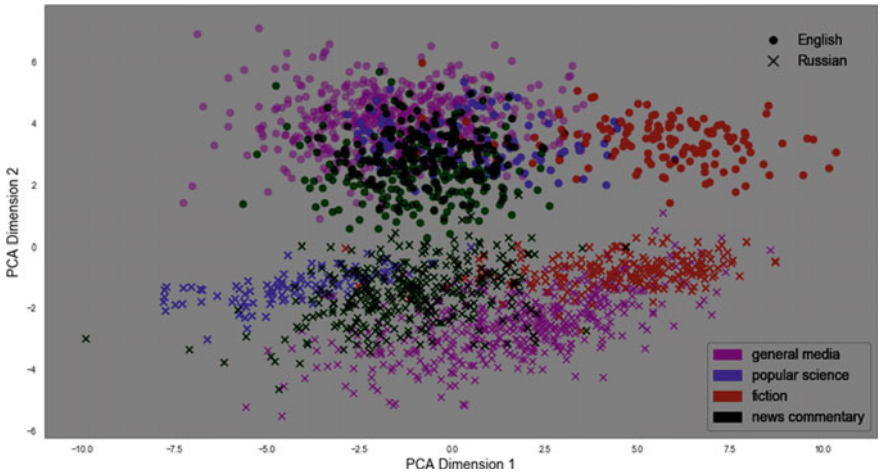

Fig. 3 PCA representation of registers in non-translations in English and Russian (56 features)

Popular science has the most expressed register differences in the cross-linguistic perspective of the four registers (notice the horizontal mismatch of the respective blue areas in the plot). *Mass-media texts* display a lot of in-category variation along the horizontal 'register' axis, especially in Russian. Judging by the upward and downward shifts of the respective clouds, this register passes some register distinctions on to Dimension 2, which ideally would capture only the language contrast. PCA on our features also struggles with distinguishing popular science and news commentary in English.

The classification results confirm that our features separate the four registers fairly well. For all 56 features, the SVM classifier, which predicted the four classes, returned 97% accuracy for each languages (F1-score 0.966 and 0.974 for English and Russian respectively). The chance level is 30% for English and 34% for Russian, with correction for imbalances between the four classes. In line with the visual impressions, most classification errors were between *mass media, commentary* and *popular science* in English and between *media* and *fiction* in Russian.

As expected in this experiment, the lexical features performed better: the 11 features were only 1% worse than 56 for English, while for Russian the decrease in performance amounted to 4%. The morphosyntactic features (45) alone were able to achieve only 78% and 81% accuracy for English and Russian, respectively. We can tentatively conclude that in our setting the register distinctions in English are conveyed through lexis to a greater extent than in Russian, where registers have more morphosyntactic specificity.

Finally, we tested whether the register distinctions in the SL are flattened out by the translation process—an assumption made by the levelling-out hypothesis (the tendency of translations to gravitate towards unmarked features in contrast to non-translated texts (Baker 1996)). The plot in Fig. 4 shows the difference in the

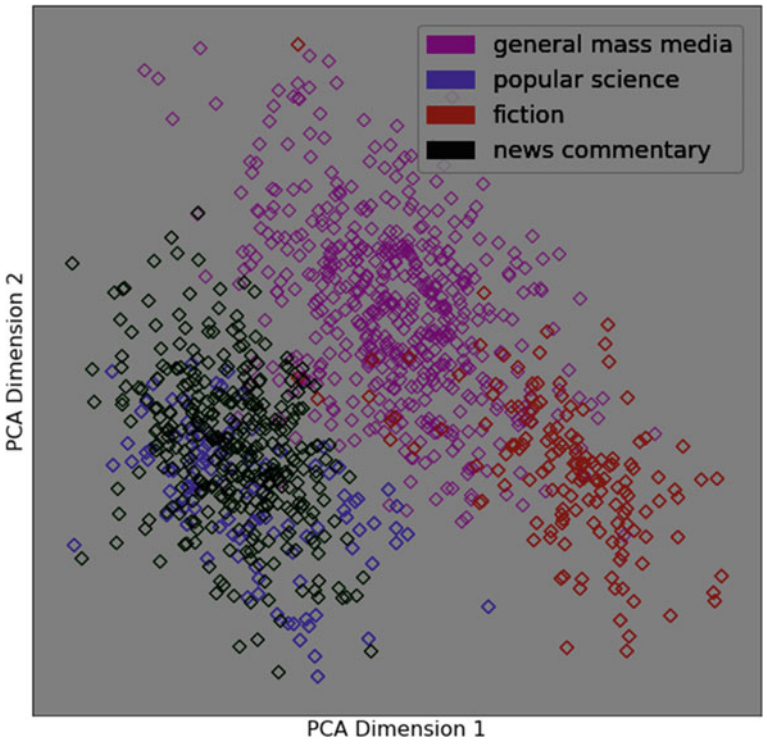

Fig. 4 Translated registers in Russian: PCA transformation of 56-dimensional feature vectors

localisation of the registers, some of which are even better separated than in the non-translated Russian (compare to the bottom part of the plot in Fig. 3). The translation process seems to import some confusion between *popular-scientific texts* and *news commentary*, on the one hand, and reinforce the separation between these two and *mass media* and *fiction*, on the other.

In this experiment, the SVM achieved the average tenfold cross-validation accuracy of 99% with a macro F1-score of 0.982 on the full feature set. Interestingly, the errors in the contingency table were between other classes than in non-translated registers: they were predictably between *news commentary* and *popular-scientific texts* (same as in the classification for English originals), rather than between *mass media* and *fiction* (as was the case in the classification for Russian originals).

Another intriguing observation is that the importance of lexical features for predicting translated registers increased compared to the texts originally written in Russian. The accuracy of register classification on the lexical feature set went up from 93 to 99% and was better than on all the 56 features. At the same time, the morphosyntax of translations introduced some noise: the classification on the 45 features from UD annotations for translation was 1% worse than for the texts

Table 4 Register distinctions in the original texts and translations for different feature sets (accuracies and macro F1 scores)

	N texts	56 features		45 morphosyntax		11 lexis	
English sources	1133	97%	0.966	78%	0.789	96%	0.955
Russian reference	1083	97%	0.974	81%	0.831	93%	0.934
Russian targets	1133	99%	0.982	80%	0.806	99%	0.983

originally written in Russian (80 vs. 81% accuracy). It indicates that the translation process does interfere with the target language register system on the structural level, but in terms of lexis translators tend to conform to the conventional distributions seen in the respective register. Table 4 systematises the results of the 4-class register classifications run on the three feature sets for each type of text in this project.

4.2 *Euclidean Distances Between Translations and Non-Translations*

To measure the apparent change of register properties in the translated language, we calculated the Euclidean distances between the register vectors for each text type (sources, targets, references). They were produced by averaging the text vectors across each category. The resulting distances are shown in Fig. 5 as a scale of the real values indicated in the diagrams. While lexical features did not contribute much to defining the specificity of translations, they were not used in measuring these distances. Besides, due to the drastic differences in the magnitudes between

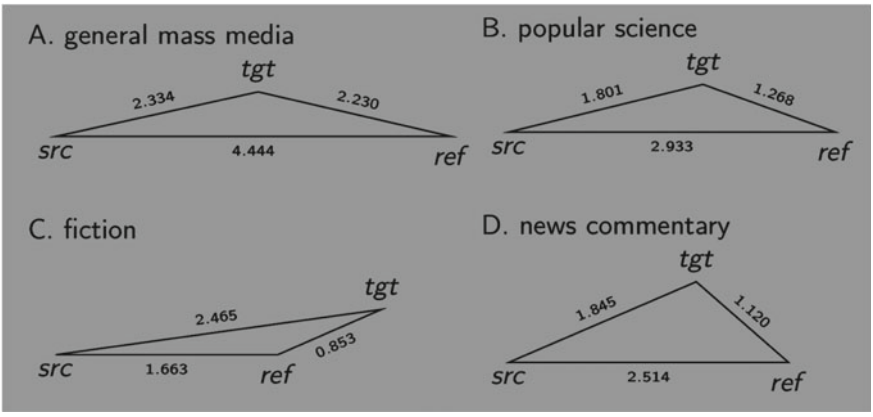

Fig. 5 Euclidean distances between the text types in each register

morphosyntactic and lexical features the latter overshadowed the former in this distance measure.

The translations in each register demonstrate some differences in how they are related to their sources and the expected target language norm. The *mass media* and *popular science* texts seem to have the most similar translationese properties, though the scale of differences is greater in the former. This generalised representation of translations from the *news commentary* subcorpora makes translations appear to be shifted more towards the TL than in the previous two registers, but at the same time the translations are more distinct from either of languages (this is indicated by the greater elevation of the *tgt* apex over the *src-tgt* plain and can be a sign of the greater amount of SL/TL-independent translationese in this register). Finally, *fiction* stands out as demonstrating an uncommon translationese shape: the diagram indicates the prevalence of adaptation or over-normalisation over shining-through effects. Note that the distances between originally-authored texts (*src* and *ref* in Fig. 5) replicate the visual results from Fig. 3.

As an additional sanity check, we computed the same measure for the random halves of the reference corpora: the average distances over 10 iterations range from 0.169 (media) to 0.712 (fiction). This confirms that translations in Russian are systematically different from the texts in the same register originally written in Russian.

The peculiarities of translationese flavours in various registers are best captured on the PCA 'register' dimension (Dimension 1) obtained from the full feature set for all texts in this project (see Fig. 6). The register properties of translations (solid coloured lines) do not necessarily replicate one language or the other, and the similarities between translations and non-translations can be seen under various register contrast conditions. The greatest mismatch of the cross-linguistic registers is seen in *general media* and *popular science*, but in the former translations tend to be in the language gap, and in the latter they appear to reproduce the TL norms. In *fiction* and *news commentary* register conventions seem to be most similar in English and Russian, and yet translations either faithfully coincide with these conventions or deviate from both.

The representations in these plots should not be taken literary, however. They do not account for the distinctions captured on the other PCA dimension and are based on the crude 2-dimensional transformation of the full feature vector. Contrary to the visual impression, translations are easily distinguishable from non-translations in all registers (Table 3).

To test Biber's claim that registers can be more distant intra-linguistically than cross-linguistically (Biber 1995: 279), we used the same approach to measure pairwise distances between registers in non-translated English and Russian. The results in Table 5, considered together with distances between *src* and *ref* for each register in Fig. 5, support this claim. In both languages fiction is more isolated from other registers structurally, especially in English, while cross-linguistically it returns the smallest distance of 1.663.

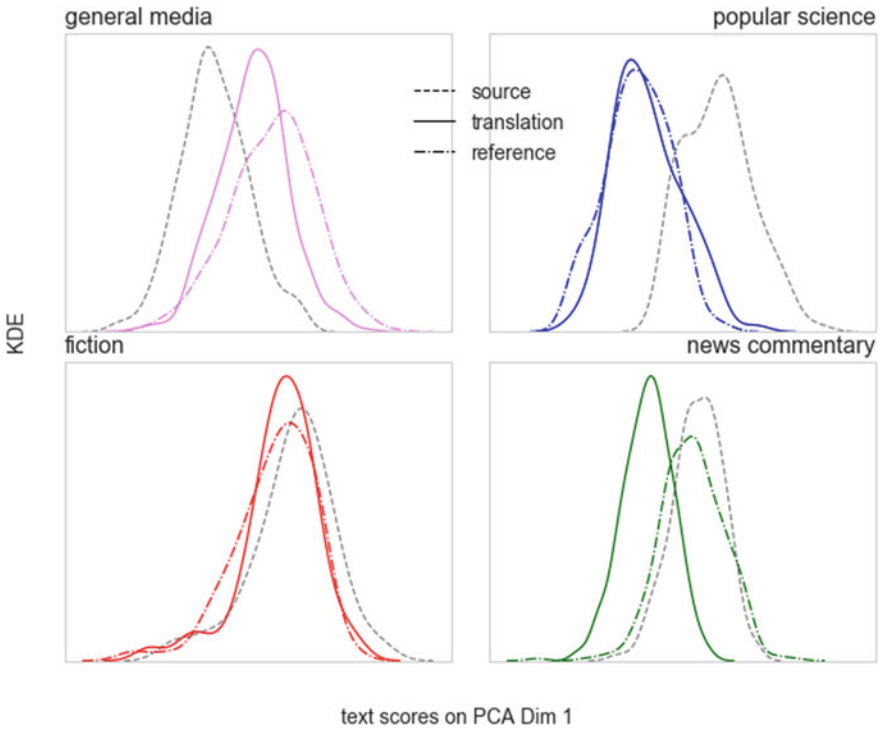

Fig. 6 Kernel Density Estimation (KDE) for the values on the PCA Dimension 1 (56 features)

Table 5 Euclidean distances between intralinguistic registers based on structural properties (values for English are under the diagonal; values for Russian are above the diagonal)

	general media	popular science	fiction	commentary
general media		1.522	3.070	2.274
popular science	0.250		4.558	0.913
fiction	5.791	5.640		5.273
commentary	0.587	0.693	6.137	

4.3 Translationese Effects and Features

In this paragraph we explore the specificity of translationese in each register through feature analysis. The results of the procedure based on the univariate analyses for *tgt-ref* (translationese), *src-ref* (language gap) and *src-tgt* (proximity to sources) are presented in Table 6. It aims to associate our features with the translationese effects described in paragraph 3.3. For the consideration of space, the table lists the 20 best translationese indicators in each register. In brackets we indicate the

Table 6 Features associated with translationese effects (based on univariate analysis of 56 features)

	shining-through	anglicisation	SL/TL independent	over-normalisation	adaptation	useless
General media	but, nnargs, relativ, interrog, whconj, neg, **parataxis**, **comp**, uoov, **trioov**, advers (23)	xcomp, **acl**, sconj (4)	epist, passives (4)	- (9)	**pverbals**, **lexdens**, **bifreq**, bioov (16)	- (0)
Popular science	**nnargs**, relativ, **deverbals**, **copula**, whconj, **aux**, **parataxis**, nsubj:pass, infs, mpred (29)	**xcomp** (2)	passives, **sconj** (5)	tempseq, simple, acl (8)	caus, advers, **mdd** (10)	**epist** (2)
Fiction	nnargs, **tempseq**, **pverbals**, whconj, **neg**, **ppron**, parataxis, **ccomp**, interrog (18)	- (0)	epist, **sconj** (3)	**xcomp**, simple, **mdd** (10)	bioov, bifreq, finites, uoov, trifreq, **lexdens** (23)	deverbals (2)
Commentary	**but**, relativ, **aux**, **parataxis**, **comp**, sup, **infs**, **advers** (20)	epist, passives, sconj (7)	ccomp (4)	**finites**, bifreq, **bioov**, **lexdens**, uoov, trioov (20)	**lexTTR**, ppron (4)	- (1)
SHARED	parataxis (6)	- (0)	- (0)	- (2)	- (0)	- (0)

Note that the three translationese features (*sconj, epist, parataxis*) shared between the lists of most important translationese indicators for our four registers, given in Table 6, are outside of the top 10 translationese indicators. Two of them (*epist, sconj*) are associated with different translationese trends in different registers

total number of features (out of 56) that fall with the respective translationese effect according to the frequency analysis. The bold font indicates the features that are among the 20 most important register contrast indicators in the respective cross-linguistic register classifications. In all four cross-linguistic register classifications (*media_src vs media_ref, fiction_src vs fiction_ref*, etc.), the accuracy on the selected features is 100%.

To identify the best translationese and the best register contrast indicators mentioned above, we relied on the Recursive Feature Elimination (RFE) algorithm in scikit-learn, a Python library. In effect, this algorithm performs an ablation study on a given feature set by recursively pruning the least important features in the multivariate setting, based an external estimator (SVR in our case). The univariate approach to feature selection based on ANOVA (SelectKBest algorithm in scikit-learn) returned a higher loss in classification performance for all experiments: on average the classification on the 20 best ANOVA features performed 2.9% worse than on the full feature set. For RFE-SVR this loss in the same experiments was only 0.9%. However, the two feature selection algorithms demonstrate contrasting performance on *popular-scientific texts*, where ANOVA is better, and on *fiction,* where the RFE 20 features do well, while ANOVA features demonstrate 5.8% decrease in performance on the F1 score. It indicates that in the first case the multivariate analysis approach fails to reveal meaningful correlations between the features frequencies, while for *fiction* the discovered patterns explain the difference between translations and non-translations better than mere univariate comparison of features. Nonetheless, the intersection between the 20 best indicators, returned by RFE and ANOVA, ranges from 9 to 13 features for different experiments.

We should reiterate here from Sect. 3.3 that 'adaptation' and 'useless' sets include features that are not translationese indicators per se, because there are no statistically significant differences for their frequencies in translations and non-translations. Nonetheless, they are not irrelevant for characterising translations. As we will see below they are also important for the machine classification.

It can be seen from Table 6 that *fiction* has the minimum number of shining-through features (18) and the maximum number of over-normalised (10) and totally adapted features (23) together, which explains the shape of the triangle for fiction shown in Fig. 5 and the matching lines in Fig. 6.

News commentary is peculiar for having the maximum number of anglicised (7) and over-normalised features (20). It makes the translated texts in this register stand out as being more distinct from both SL and TL, indicated in Fig. 5 as a greater elevation of the translations apex over the *src-ref* plain and in Fig. 6 by the location of the translations outside the area shared by sources and reference.

Another immediate observation is that the registers tend to have no shared features for the suggested translationese effects, except shining-through and over-normalisation. However, even these effects seem to be achieved through widely different sets of features: only 6 features are shared among the average of 23 features for shining-through (*nnargs, relativ, whconj, parataxis, interrog, mpred*) and there are two shared over-normalisation indicators (*possdet, correl*).

It is also clear from Table 6 that, in terms of the number of features, shining-through is by far the most important type of deviation from the expected norm in translation.

We failed to detect any pattern in the relation of the features prominent in cross-linguistic register classifications (in bold) and the features important for the translationese classification (named in Table 6). Some of the contrastive register features are adapted to the TL norms and some are carried over from the SL.

The lists in Table 6 should be taken with caution, though. One limitation is that some features have negligibly small values and calculations for them are less reliable. For others, the differences in frequencies can be significant but the effect size is small. Besides, the impact of some feature sets associated with a given translationese effect can be comparatively small in the classification task, despite their size.

To verify the observations from the univariate analysis, we extracted the absolute weights of the features associated with each effect for each register from the SVM translationese classifier, and calculated the mean and standard deviation (SD) for these weights. Feature weights from a linear SVM classifier can be used to identify the features that contributed most to the classifier decision. This approach is known to be reliable in feature ranking (Chang and Lin 2008) . Additionally, we looked at the effect size (measured as Cohen's d) for the features with significant differences in frequencies between translations and non-translations (at $p < 0.05$). We report the findings for the most prominent trends by register in Table 7.

It can be seen from Table 7 that the effect size in the last column did not correlate with the classifier weights. Some features with the observed greater magnitude of

Table 7 The most prominent translationese effects in each register (in the order of importance based on the classifier weights)

	effect	N features	Mean weights	SD	Cohen's d
General media	anglicisation	4	0.645	0.210	0.851
	shining-through	23	0.348	0.395	0.232
	adaptation	16	0.325	0.316	–
Popular science	SL/ TL-independent	5	0.243	0.137	0.063
	adaptation	16	0.223	0.118	–
	shining-through	29	0.183	0.161	0.079
Fiction	shining-through	18	0.483	0.371	0.274
	over-normalisation	10	0.451	0.387	0.099
	adaptation	23	0.398	0.242	–
News commentary	adaptation	4	0.662	0.280	–
	anglicisation	7	0.583	0.380	0.601
	shining-through	20	0.553	0.302	0.361
	over-normalisation	20	0.449	0.292	0.200

differences were not selected by the algorithm as important. The comparison of the performance of the two feature selection algorithms, given above, shows that from a machine point of view finding patterns in the data is more effective than relying on separate features in most cases. It is not clear, however, which translationese effects are more visible (if any) to a human user.

5 Register-Based Translationese Varieties

We have seen that professional translations deviate from non-translations in the TL in all registers, which is particularly noticeable on the structural level. These deviations accommodate a number of trends, including shining-through, over-normalisation and adaptation.

The size and the combination of the translationese effects is register-specific, especially if we consider the associated sets of features. Our registers have just one intersecting translationese indicator in the top 20 most important translationese features (*parataxis*). It captures one strong and universal trend across our registers in translations—to spot more introductory and parenthetical elements and non-linear syntax. In general, the lexical features perform much worse than the structural (morphosyntactic) ones, with the difference in accuracies of the translationese classifications ranging from 22% (*popular science*) to 5% (*news commentary*).

As for the translationese effects, shining-through is the strongest trend in all registers, judging by the number of features identified as such and by their weights in the classifier. It is complemented by tendencies with less features, but sometimes higher prominence, to create a unique linguistic make-up for each category, described below.

1. In *general media* the strong pull towards the SL is emphasised by anglicised features and is to an extent counter-balanced by the fully adapted features. The prevailing trend is still to exploit the SL patterns where possible. On the one hand, it is understandably hard for translators to assimilate the considerable cross-linguistic distance in this register. On the other hand, the expected TL norm is less defined in Russian mass-media corpus than in the other registers (note the broad spread of the media texts in Russian in Fig. 3).
2. *Popular scientific* translations have the record number of shining-through indicators, but a third of them are lexical features that do not contribute much to the translationese classification according to the classifier weights and the analysis above, particularly in this register. The prevailing trend is towards adaptation, which is reasonable, if we bear in mind a clearer delineation of this register in the TL. This is the only register where the SL/TL-independent translationese features are important for the classifier. Notably, this register has a significantly lower frequency of passives and significantly higher frequency of

subordinate conjunctions than in either original English or Russian, without a cross-linguistic contrast for this feature.

3. *Fiction* has the least shining-through indicators, and yet, according to the classifier, these features rank high in importance. The second strongest tendency is over-normalisation (or russification). The pull towards the TL norm is reinforced by the considerable input from the record number (23) of fully adapted features. This register appears to be the most Russian-like in translation.
4. In *news commentary* the few fully adapted features are assigned the biggest weights. We will highlight that this register has the largest list of over-normalised features (20) with relatively high weights. The other two effects with comparably high average feature weights are anglicisation and shining-through. It looks like this register is sharply torn between the two languages.

The suggested feature sets are also fairly reliable for defining the contrastive properties of the registers. They can be used to distinguish the four text categories with 97% accuracy. However, the importance of morphosyntactic and lexical features is reverse compared to the translationese classification. The lexical features outperformed morphosyntax in register classification. Besides, we were able to capture less morphosyntactic variation across English registers than across their Russian counterparts. The translated registers exhibit clearer register distinctions than the comparable TL non-translations, especially on the lexical level. However, using morphosyntactic features only, it is more difficult to predict registers in translations than in non-translations. It means that on the structural level the translated registers are a bit less well-defined than non-translations in the TL (see Table 4). It indicates that the translation process does not level out the distinctions between the registers. Additionally, one can claim that the register conversions are exaggerated and amplified, which leads to (1) higher similarity of translated texts from one register and/or to (2) greater distances between the registers.

We put these two hypotheses to a quick test by (1) comparing the averaged distance from centroid (corpus average vector) to each text vector for translated and non-translated registers in Russian ('degree of homogeneity' measure) and by (2) measuring the Euclidean distances between the translated registers (and use the distances in Table 5 for reference).

These experiments show that (1) translations are less diverse than their non-translated counterparts in all registers; (2) the second hypothesis holds only for translated fiction, which is even stronger isolated from the other registers than in non-translations (see Fig. 4), but not for the other registers, where the relatively clear distinctions in the original Russian are blurred in translation in terms of morphosyntax.

Now, the question is whether the amount and type of translationese can be explained by the degree of the cross-linguistic similarity between the registers or they have to be attributed to the extralinguistic factors such as translational norms operating in the contemporary professional community and the other translation process variables such as the input of editors and working conditions. Or in other

words, is translationese a function of the linguistic distance between registers? From our observations in Fig. 6 this not likely to be the case.

The previous research on human translations reports different results in this respect based on translationese properties induced by different SLs. Diana Santos observes that languages closeness as a factor in translations has a paradoxical effect: 'the closer the languages the larger the quantity of false friends and cognates, both in lexicon and in grammar', because it is easier to carry over the SL properties (Santos 1995: 64). Sominsky and Wintner concluded that 'translationese is more pronounced, and interference is more powerful, when the two languages are more distant' based on their classification result in the SL detection task (Sominsky and Wintner 2019: 1138).

An apparent reconciliation for these competing observations is found in (Nikolaev et al. 2020). They explore the predictability of translations and find differences between translations from structurally similar and structurally dissimilar source languages. In the former case translations tend to employ an intersection of syntactic patterns found in both languages, which makes them less rich, more repetitive, in the latter case 'translators find it hard to fully rework the original morphosyntactic patterns and produce unpredictable/entropic non-idiomatic translations' (Nikolaev et al. 2020).

In our setting this should be observed as the difference for the degrees of homogeneity of the respective translated corpora: the more cross-linguistically similar registers (fiction and news commentary) should demonstrate higher degree of homogeneity in translation. This was indeed observed in our data where the averaged vector distance to centroid was 3.050 and 2.488 for *fiction* and *news commentary*. For more distant registers—*media* and *popular science*—this measure returned 3.354 and 3.281. Note that for distances the smaller numbers mean more similar texts.

6 Conclusion

In this chapter we investigated the impact of register on the properties of translations in the English-Russian language pair. We used parallel corpora of professional translations and comparable reference corpora from the national corpora in four registers (general media, popular science, fiction, news commentary) to explore the relations of the original texts in the two languages and the translated registers. Our approach exploits linguistically interpretable features and is contingent on their selection and effectiveness for capturing differences between registers, on the one hand, and translationally relevant text types (sources, targets, and TL reference), on the other. For both tasks we tested and described the behaviour of 45 morphosyntactic and 11 lexical features. The former represent the text structure in terms of general text properties, frequencies of PoS and syntactic phenomena, the latter provide text characteristics from the point of view of lexical predictability scores and the ratios of high-frequency and low-frequency n-grams.

The results demonstrate that our experimental setup, including the suggested features, is reliable for distinguishing registers in translated and non-translated language as well as for predicting translations in each register, and, therefore, can be used for revealing the register-related specificity of translations in the given language pair. Admittedly, the features used are language pair specific, and out findings apply for English-to-Russian translation. We leave testing the suggested methodology on other language pairs for future work.

Our findings contribute to the understanding of the linguistic properties of Russian translations from English in general and to the investigation of their specificity across registers. We suggested a distance-based method to estimate the general shapes of translationese in a register-balanced corpus for comparative analysis, taking into account the cross-linguistic properties of each register. A novel bottom-up approach was used to associate the linguistic features with a number of translationese effects and to disentangle the opposite translational tendencies.

We demonstrated that (1) professional translations in all registers are easily distinguishable from non-translations and these distinctions mostly involve morphosyntactic, rather than lexical, properties; (2) more than a third of all translationese indicators have their frequencies shifted towards the values observed in the SL (shining-through features), but their actual impact on the classification results varies and can be overshadowed by strong features representing other trends; (3) each register generates a unique form of translationese, with the various translationese effects contributing to a different extent and being realised through widely diverging sets of features; (4) translated registers have more regularity in feature frequencies and higher intra-category homogeneity than their non-translated English and Russian counterparts. The more cross-linguistically similar registers seem to generate the more homogeneous translations.

One important message from this research is that human translations vary depending on the register. Some of this variation can be explained linguistically. However, some of the translation strategies are likely to be dictated by the established practice and professional norms operating in each register, including the tolerance to translationese.

The scope of this work did not allow us to perform in-depth analysis of the individual features that were identified as having translationally interesting behaviours. The machine learning results can be convincing mathematically, but they remain a noumenon unless they are related to human perception.

Although this research takes into account the specificity of the given language pair, it would certainly be interesting to extend it to other target languages or language pairs. The more immediate development would be to consider other registers in the explored language pair, if the necessary corpus resources are available. We hope that this research will promote the idea that register is one of the central factors in translationese studies, even if its impact on the translation properties is not defined by purely linguistic matters.

Acknowledgements The research presented in this paper has been partially carried out in the framework of projects in the framework of the projects VIP (FFI2016-75831-P), TRIAGE (UMA18-FEDERJA-067) and MI4ALL (CEI-RIS3). The authors would like to thank two anonymous reviewers for their valuable comments.

Appendix

The UD-based and list-based features in alphabetical order.

Preliminary Notes

1. Normalisation measures

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 types of discourse markers, negative particles, passives, types of verb forms, relative clauses, correlative constructions, adverbial clauses introduced by pronominal adverbs coordinating and subordinating conjunctions, simple sentences, number of clauses per sentence, are normalised to the number of sentences (30 features). Such features as personal, possessive 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 have 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.

2. Groups of discourse markers

The classification of connectives (discourse markers) follows the descriptions in Halliday and Hasan (1976) and in Biber et al. (1999). Table A has the number of items in each group and most frequent examples. The lists were initially produced independently from grammar reference books, dictionaries of function words and relevant research papers (for English we used Biber et al. (1999), Fraser (2006), Liu (2008); for Russian—Novikova (2008), Priyatkina (2015), Russian Grammar (Shvedova 1980) to name just a few sources for each language). After the initial selection, the lists were verified for comparability. Following Fraser (2006), discourse markers are treated functionally and include items of various morphological and structural types (conjunctions, adverbs, particles, parenthetical phrases). Though most items on the lists are set phrases, we allowed for possible lexical and structural variability at the extraction time. We also used orthography and punctuation to disambiguate our items. The output of the extraction procedure was manually checked to exclude greedy matching.

Table 8 Number of listed connectives and discourse markers by category for each of the project languages and top five most frequent items

	English	Russian
Additive	52	52
	Also, such as, for example, not only, for instance, in particular, moreover, in other words, namely	Также, при этом, например, кроме того, в частности, к тому же, на самом деле, а именно, иными словами, точнее, причем, вдобавок
Adversative	46	34
	Still, however, rather than, instead, though, on the other hand, in fact, despite	Однако, хотя, впрочем, правда, несмотря на, в отличие от, вместе с тем, всё-таки, но на самом деле, наоборот, напротив, зато
Causative	42	49
	Because, so, due to, so that, therefore, as a result, after all, for this reason, consequently	Потому, поэтому, поскольку, ведь, так,, в результате, ради того, чтобы, затем, что, получается, в этом случае, в связи с тем, дабы, тем более что
Temporal and sequential	110	48
	While, since, soon, and then, eventually, further, anyway, thus, at the same time, ultimately, meanwhile	Пока, наконец, затем, в целом, в то время, как, в заключение, в конце концов, во-первых, в то же время
Epistemic markers	64	86
	Really, at least, perhaps, of course, probably, in any case, for sure, in reality, no doubt, arguably, clearly, indeed, I/we think, I/we am/are (un) convinced/sure	Конечно, возможно, может быть, действительно, говорят, на мой взгляд, якобы, полагаю, по сути, в любом случае, кажется, бесспорно, пожалуй

3. The alphabetic list of 45 morphosyntactic features

acl
finite and non-finite clausal modifier of noun (adjectival clause), including relative clauses as a subtype (used only in EN and RU); extraction is based on UD default annotation (e.g. *the person showing (acl) her around; help people do something to overcome (acl) it; людей, следящих (acl) за политикой)*

addit
additive connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

advers
adversative (contrastive) connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

attrib
adjectives and participles functioning as attributes; all words tagged as ADJ or VerbForm = Part with the *amod* dependency to their head (e.g. *the rising sun; the coloured face; fried green tomatoes)*

aux
auxiliary verbs; extraction is based on UD default annotation

aux:pass
auxiliary verbs in passive forms; extraction is based on UD default annotation

but
contrastive coordinating conjunction *but* (*но*), if not followed but *also/и*, *также* and not in the absolute sentence end

caus
causative connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

ccomp
clausal complement as annotated in UD (e.g. *help people to do (ccomp) smth; не ожидали, что придет (ccomp))*

cconj
coordinating conjunctions: lemmas in *and, or, both, yet, either, &,* nor*, plus, neither, ether / и, а, или, ни, да, причем, либо, зато, иначе, только, ан, и/или, иль* tagged CCONJ. Lists are used to filter out noise.

comp
comparative degree of comparison for adjectives and adverbs; synthetic forms are extracted based on the tag Degree = Comp, while analytical forms are counted as adjectives and adverbs with a dependent *more/более* (*больший)*

copula
copula verbs; lemmas of *be, быть, это* that have a *cop* relation to their head, excluding constructions with *there* as head for English

correl
correlative constructions of all types, where a PRON/DET (*those, such*) is syntactically or semantically connected to subsequent CONJ. In English they make a subset of relative clauses; in Russian they can also be a subtype of a clausal complement (e.g. *of those who voted for him, raising the living standards of those that are poor)*

demdets
pronominal determiners; lemmas in the function *det* from the lists *this, some, these, that, any, all, every, another, each, those, either, such* / *этот, весь, тот, такой, какой, каждый, любой, некоторый, какой-то, один, сей, это, всякий, некий, какой-либо, какой-нибудь, кое-какой*

deverbals
deverbal nouns, names of processes, actions, states. The extraction for English accounts for affixation (with most productive *-ment, -tion/ -ing, -tion*) and conversion as types of derivation. In the first case the output is filtered with an empirically driven stop list. Converted nouns are counted from a list of true procedural nouns that were not fully substantivised. To produce this list we looked through the nounal occurrences of lemmas that also appear as verbs and filtered out items that prevail in their fully substantivised lexico-semantic variants in our data (such as *design, set, measure, mark, press, stick, cross, trap, handle*). For Russian we extracted nouns in *-тие, -ение, -ание, -ство, -ция, -ота* and employed a 150-items long stop list to exclude fully substantivised words such as *собрание, месторождение, министерство, телевидение, творчество, решение*.

epist
epistemic stance discourse markers; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

finites
verbs in finite form; extraction is based on UD default annotation VerbForm = Fin

indef
noun substitutes, i.e. pronouns par excellence, of indefinite, total and negative semantic subtypes; extraction is based on PRON tag with a filter list: *anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, somebody, someone, something, elsewhere, nowhere, everywhere, somewhere, anywhere* / *когда, где, куда, откуда, отчего, почему, зачем* and words with *-то*|*-нибудь*|*-либо*, except starting with *какой*; and items from *кто-кто, кого-кого, кому-кому, кем-кем, ком-ком, что-что, чего-чего, чему-чему, чем-чем, куда-куда, где-где*

infs
infinitives: all cases of a verb form tagged VerbForm = Inf with a dependent *to* particle and cases of true bare infinitive, excluding after modal verbs and *have to*, *going to* and modal adjectival predicates, but including cases after *help, make, bid, let, see, hear, watch, dare, feel*. For Russian all occurrences of verb forms with the feature VerbForm = Inf except after modal predicates and with the dependent *быть* to exclude future forms (e.g. *отношения будут ухудшаться*).

interrog
interrogative sentences: all sentences ending in ?

lexdens

lexical density: ratio of PoS disambiguated content words types (look_VERB vs look_NOUN) to all tokens

lexTTR
lexical type-to-token ratio: ratio of PoS disambiguated content words types (look_VERB vs look_NOUN) to their tokens. Content words include lemmas in ADJ, ADV, VERB, NOUN part-of-speech categories.

mdd
mean dependency distance (MDD, aka comprehension difficulty) as 'the distance between words and their parents, measured in terms of intervening words' (Jing and Liu 2015: 162)

mhd
mean hierarchical distance (MHD, aka production (speaker's difficulty) as the average value of all path lengths travelling from the root to all nodes along the dependency edges (Jing and Liu 2015: 164)

mpred
modal predicates; for English all verbs tagged as MD in XPOS, except *will/shall*, constructions with have-to-Inf and all adjectival modal predicates (given a list of 17 predicatives such as *impossible, likely, sure* with a dependent AUX). For Russian: lemma *мочь*, lemma *следовать* with a dependent infinitive, three modal adverbs (*можно, нельзя, надо*) and 11 adjectives from the modal predicative list in the short form Variant = Short (e.g. *должен, способный, возможный*)

mquantif
adverbial quantifiers; listed lemmas tagged ADV. The support lists include 37 English items (e.g. *barely, completely, intensely, almost*), 80 Russian items (*абсолютно, полностью, сплошь, необыкновенно, достаточно, совершенно, невыносимо, примерно*). For Russian we additionally provide for functionally similar non-adverbial quantifiers such as *еле, очень, вшестеро, невыразимо, излишне, еле-еле, чуть-чуть, едва-едва, только, капельку, чуточку, едва*.

neg
negative particles or main sentence negation: counts of lemmas in *no, not, neither* / нет, не

nnargs
core verbal arguments represented by nouns or proper names; ratio of nouns and proper names in the functions of *nsubj, obj, iobj* to the count of these functions

nsubj:pass
subjects of verbs in the passive voice; extraction is based on UD default *nsubj:pass* annotation

numcls
number of clauses per sentence; number of relations from the list *csubj, acl:relcl, advcl, acl, xcomp, parataxis* annotated in one sentence

passives
passive constructions with expressed agentive role; all verbs tagged Voice = Pass and a dependent aux:pass (for English). For Russian we account for two morphological forms (*война велась, политика была направлена*) and for semantic passive (*стадион возводят на новом месте, во Владикавказе ему готовят радушную встречу*)

parataxis
asyndatically connected coordinated clauses (often direct speech or clauses joined ':' or a ';' as well as parenthetical clauses); extraction is based on UD default annotation

pasttense
verbs in the past tense: all occurrences of the feature Tense = Past

pied
correlative constructions with displaced (pied-piped) preposition (e.g. *technology for which Sony could take credit; speech in which he made this argument; о таком, о каком вы не слыхали; скандал, в котором; трагедии, с которыми, в той конструкции, в какой она*)

possdet
possessive pronouns; for English lemma in *my, your, his, her, its, our, their* tagged DET, PRON and Poss = Yes. For Russian lemma in *мой, твой, ваш, его, ее, её, наш, их, ихний, свой* tagged DET

ppron
personal pronouns; tokens tagged PRON, with any value of attribute Person = that do not have Poss = Yes feature and are on the list: *i, you, he, she, it, we, they, me, him, her, us, them* / *я, ты, вы, он, она, оно, мы, они, меня, тебя, его, её, ее, нас, вас, их, неё, нее, него, них, мне, тебе, ей, ему, нам, вам, им, ней, нему, ним, меня, тебя, него, мной, мною, тобой, тобою, Вами, им, ей, ею, нами, вами, ими, ним, нем, нём, ней, нею*

pverbals
participles: for English all occurrences of VerbForm = Part or VerbForm = Ger not in attributive function *amod* or part of an analytical form. For Russian VerbForm = Part not in the short form and not in the attributive function, without a dependent auxiliary, and VerbForm = Conv without dependent auxiliary (e.g. a*fter years of translating emails, webinars and other materials*)

relativ
all relative clauses, including correlative constructions and pied-piping construction. Extraction is based on affirmative sentences only. For English: *which, that, whose, whom, what, who* tagged as PRON, excluding cases when relative PRON has a dependent preposition and follows its head (e.g. *But we will return to that (PRON) later*). For Russian: *который, что, кто, какой* and a comma in the left window of 3

sconj
subordinating conjunctions: lemma in *that, if, as, of, while, because, by, for, to, than, whether, in, about, before, after, on, with, from, like, although, though, since, once, so, at, without, until, into, despite, unless, whereas, over, upon, whilst, beyond, towards, toward, but, except, cause, together / что, как, если, чтобы, то, когда, чем, хотя, поскольку, пока, тем, ведь, нежели, ибо, пусть, будто, словно, дабы,раз, насколько, тот, коли, коль, хоть, разве, сколь,ежели, покуда, постольку* tagged SCONJ. Lists are used to filter out noise.

sentlength
number of words per sentence averaged over all sentences in the text. The extraction accounts for typical sentence tokenisation errors such as sentences ending in:,;, Mr., Dr.

simple
simple sentence; a sentence where no words have relations: *csubj, acl:relcl, advcl, acl, xcomp, parataxis*

sup
superlative degree of comparison for adjective and adverbs; synthetic forms are extracted based on the tag Degree = Sup, while analytical forms are counted as adjectives and adverbs with a dependent most/наиболее/самый and for Russian words starting with *наи-* with the exception of a few homonymous adverbs (*наискосок*)

tempseq
temporal and sequential connectives; cumulative frequency of the list items normalised to the number of sentences; see description in Table A

whconj
adverbial clause introduced by a pronominal ADV *when, where, why / когда, где, куда, откуда, отчего, почему, зачем*

xcomp
a predicative or clausal complement without its own subject, annotated after phrasal verbs (e.g. *started to sing*), in case of infinitive constructions (e.g. *asked me to leave*), etc.; extraction is based on UD default annotation

References

Aharoni, R., M. Koppel, and Y. Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL 2014)*, Vol. 1: Long Papers, ed. K. Toutanova, and H. Wu, 289–295. Association for Computational Linguistics https://doi.org/10.3115/v1/p14-2048.

Arase, Y., and M. Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st annual meeting of the association for computational linguistics*, Vol. 1: Long Papers, ed. H. Schütze, F. Pascale, and M. Poesio, 1597–1607. Association for Computational Linguistics.

Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and technology: In honour of John Sinclair*, ed. M. Baker, G. Francis, and E. Tognini-Bonelli, 232–250. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.64.15bak.

Baker, M. 1996. Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and translation: Studies in language engineering, in honour of Juan C. Sager*, ed. H. Somers, 175–186. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/btl.18.17bak.

Baroni, M., and S. 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. https://doi.org/10.1093/llc/fqi039.

Becher, V. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts* [Doctoral dissertation, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky]. https://ediss.sub.uni-hamburg.de/bitstream/ediss/4186/1/Dissertation.pdf.

Biber, D. 1988. *Variation across speech and writing*, 2nd ed. Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519871.

Biber, D., and S. Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and R. Quirk. 1999. *Longman grammar of spoken and written English*, vol. 2. Cambridge, MA: The MIT Press.

Castagnoli, S. 2009. *Regularities and variations in learner translations: a corpus-based study of conjunctive explicitation* [Doctoral dissertation, University of Pisa, Italy]. ETD System, electronic theses and dissertations repository. https://etd.adm.unipi.it/t/etd-04252009-135411/.

Castagnoli, S., D. Ciobanu, K. Kunz, N. Kübler, and A. Volanschi. 2011. Designing a learner translator corpus for training purposes. In *Corpora, language, teaching, and resources: From theory to practice*, Vol. 12, ed. N. Kubler, 221–248. Frankfurt: Peter Lang.

Chang, Y., and C. Lin. 2008. Feature ranking using linear SVM. In *Proceedings of the workshop on the causation and prediction challenge at WCCI 2008*, ed. I. Guyon, C. Aliferis, and G. Cooper, 53–64. Proceedings of Machine Learning Research.

Corpas Pastor, G. 2008. *Investigar con corpus en traducción: Los retos de un nuevo paradigma*. Frankfurt: Peter Lang. https://doi.org/10.4000/bulletinhispanique.1301.

Corpas Pastor, G., R. Mitkov, N. Afzal, and V. Pekar. 2008. Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the 8th conference of the association for machine translation in the Americas (AMTA'08)*, 21–25.

Delaere, I. 2015. *Do translations walk the line? Visually exploring translated and non-translated texts in search of norm conformity*. [Doctoral dissertation, Ghent University]. Academic Bibliography. https://biblio.ugent.be/publication/5888594.

Dipper, S., M. Seiss, and H. Zinsmeister. 2012. The use of parallel and comparable data for analysis of abstract anaphora in German and English. In *Proceedings of the 8th international*

conference on language resources and evaluation (LREC 2012), ed. N. Calzolari, Kh. Choukri, Th. Declerck, M. Uğur Doğan, et al., 138–145. European Language Resources Association.

Diwersy, S., S. Evert, and S. Neumann. 2014. A semi-supervised multivariate approach to the study of language variation. In *Linguistic variation in text and speech, within and across languages*, ed. B. Szmrecsanyi, and B. Wälchli, 174–204. Berlin: De Gruyter Mouton.

Duff, A. 1981. *The third language: Recurrent problems of translation into English*. Oxford: Pergamon.

Eetemadi, S., and K. Toutanova. 2015. Detecting translation direction: A cross-domain study. In *Proceedings of NAACL-HLT 2015 student research workshop (SRW)*, ed. D. Inkpen, S. Muresan, Sh. Lahiri, K. Mazidi, and A. Zhila, 103–109. https://doi.org/10.3115/v1/N15-2014.

Evert, S., and S. Neumann. 2017. The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In *Empirical translation studies: New methodological and theoretical traditions*, vol. 300, ed. G. De Sutter, M. Lefer, and I. Delaere, 47–80. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110459586-003.

Fraser, B. 2006. Towards a theory of discourse markers. In *Approaches to discourse particles*, ed. K. Fischer, 189–204. London: Elsevier.

Frawley, W. 1984. Prolegomenon to a theory of translation. In *Translation: Literary, linguistic & philosophical perspectives*, ed. W. Frawley, 159–175. Newark: University of Delaware Press.

Gellerstam, M. 1986. Translationese in Swedish novels translated from English. In *Translation studies in Scandinavia*, ed. L. Wollin and H. Lindquist, 88–95. Lund: CWK Gleerup.

Goutte, C., D. Kurokawa, and P. Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of the 12th machine translation summit (MT Summit XII)*, 81–88.

Graham, Y., B. Haddow, and P. Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 72–81). Association for Computational Linguistics.

Halliday, M.A.K., and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Halliday, M., and R. Hasan. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective* (2nd ed.). Oxford University Press.

Hansen-Schirra, S. 2011. Between normalization and shining-through. Specific properties of English-German translations and their influence on the target language. In *Multilingual discourse production: Diachronic and synchronic perspectives*, ed. S. Kranich, 133–162. Amsterdam: John Benjamins.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 sixth workshop on statistical machine translation*, ed. Ch. Callison-Burch, Ph. Koehn, Ch. Monz, and O. Zaidan, 187–197. Association for Computational Linguistics.

Ilisei, I., D. Inkpen, G. Corpas Pastor, and R. Mitkov. 2010. Identification of translationese: A machine learning approach. *International conference on intelligent text processing and computational linguistics*, 503–511.

Jiang, Z., and Y. Tao. 2017. Translation universals of discourse markers in Russian-to-Chinese academic texts: A corpus-based approach. *Zeitschrift Fur Slawistik* 62 (4): 583–605. https://doi.org/10.1515/slaw-2017-0037.

Jing, Y., and H. Liu. 2015. Mean hierarchical distance augmenting mean dependency distance. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, ed. J. Nivre and E. Hajicova, 161–170. Uppsala University.

Karakanta, A., and E. Teich. 2019. Detecting and analysing translationese with probabilistic language models translationese. In *Translation in Transition* 4: 38–39.

Katinskaya, A., and S. Sharoff. 2015. Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. In *Proceedings of the 5th workshop on Balto-Slavic natural language processing*, ed. J. Piskorski, L. Pivovarova, J. Šnajder, H. Tanev, and R. Yangarber, 65–74. INCOMA Ltd. http://www.aclweb.org/anthology/W15-5311.

Koppel, M., and N. Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, Vol.

1, ed. D. Lin, Yu. Matsumoto, and R. Mihalcea, 1318–1326. Association for Computational Linguistics.

Kruger, H., and B. Rooy. 2012. Register and the features of translated language. *Across Languages and Cultures* 13 (1): 33–65. https://doi.org/10.1556/Acr.13.2012.1.3.

Kruger, H., and B. van Rooy. 2010. The features of non-literary translated language: A pilot study. In *Proceedings of using corpora in contrastive and translation studies (UCCTS 2010)*, ed. R. Xiao, 59–79.

Kunilovskaya, M. 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"*, Vol. 2, ed. V.P. Selegey, A.V. Baytin, V.I. Belikov, I.M. Boguslavsky, B.V. Dobrov, et al., 221–233. Computational Linguistics and Intellectual Technologies.

Kunilovskaya, M., and A. Kutuzov. 2018. Universal dependencies-based syntactic features in detecting human translation varieties. In *Proceedings of the 16th international workshop on treebanks and linguistic theories (TLT16),* ed. J. Hajič, 27–36. Association for Computational Linguistics.

Kunilovskaya, M., and E. Lapshinova-Koltunski. 2020. Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of the 12th conference on language resources and evaluation (LREC 2020)*, ed. N. Calzolari, F. Bechet, Ph. Blache, Kh. Choukri, et al., 4102–4112. The European Language Resources Association (ELRA).

Kunilovskaya, M., and E. Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the 2nd workshop on human-informed translation and interpreting technology (HiT-IT 2019)*, ed. I. Temnikova, C. Orasan, G. Corpas Pastor, and R. Mitkov, 47–56. INCOMA Ltd. https://doi.org/10.26615/issn.2683-0078.2019_006.

Kutuzov, A., and M. Kunilovskaya. 2014. Russian learner translator corpus: Design, research potential and applications. In *Proceedings of the 17th international conference text, speech and dialogue*, vol. 8655, ed. P. Sojka, A. Horák, I. Kopeček, and K. Pala, 315–323. Springer.

Lapshinova-Koltunski, E. 2017. Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. In *Empirical translation studies. New theoretical and methodological traditions*, vol. 300, ed. G. De Sutter, M. Lefer, and I. Delaere, 207–234. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110459586-008.

Lapshinova-Koltunski, E., and M. Zampieri. 2018. Linguistic features of genre and method variation in translation: A computational perspective. *The grammar of genres and styles: From discrete to non-discrete units*, (TiLSM, *320*), 92–112. Berlin: De Gruyter Mouton.

Lee, D.Y.W. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5 (3): 37–72. https://doi.org/10.1016/S1364-6613(00)01594-1.

Lembersky, G., N. Ordan, and S. Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38 (4): 799–825. https://doi.org/10.1162/COLI_a_00111.

Lijffijt, J., T. Nevalainen, T. Säily, P. Papapetrou, K. Puolamäki, and H. Mannila. 2016. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* 31 (2): 374–397. https://doi.org/10.1093/llc/fqu064.

Liu, D. 2008. Linking adverbials: An across-register corpus study and its implications. *International Journal of Corpus Linguistics* 13 (4): 491–518. https://doi.org/10.1075/ijcl.13.4.05liu.

Martin, J.R. 1992. *English text: System and structure*. Amsterdam: John Benjamins.

Nakamura, S. 2007. Comparison of features of texts translated by professional and learner translators. In *Proceedings of the 4th corpus linguistics conference.* University of Birmingham.

Neumann, S. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: De Gruyter Mouton.

Nikolaev, D., T. Karidi, N. Kenneth, V. Mitnik, L. Saeboe, and O. Abend. 2020. Morphosyntactic predictability of translationese. *Linguistics Vanguard*, 6 (1).

Nini, A. 2019. The multi-dimensional analysis tagger. In *Multi-dimensional analysis: research methods and current issues*, ed. T. Berber Sardinha, and M. Veirano Pinto, 67–94. London; New York: Bloomsbury Academic. https://doi.org/10.5040/9781350023857.0012.

Nisioi, S., and L.P. Dinu. 2013. A clustering approach for translationese identification. In *Proceedings of the international conference recent advances in natural language processing (RANLP 2013),* ed. R. Mitkov, G. Angelova, and K. Bontcheva, 532–538. INCOMA Ltd. http://www.aclweb.org/anthology/R13-1070.

Novikova, N.I. 2008. Connectives as cohesive devices in an asyndetic composite sentence [Konnektory kak svjazujushhie sredstva v bessojuznom slozhnom predlozhenii]. In Herald of the Voronezh state Architecture University, advanced linguistic and pedagogical research series [Ser.: Sovremennye lingvisticheskie i metodiko-didakticheskie issledovanija], 92–100.

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

Popescu, M. 2011. Studying translationese at the character level. In *Proceedings of the international conference recent advances in natural language processing (RANLP 2011)*, 634–639. http://aclweb.org/anthology/R11-1091.

Popovic, M. 2020. On the differences between human translations. In *Proceedings of the 22nd annual conference of the European association for machine translation*, ed. A. Martins, H. Moniz, S. Fumega, M. Martins, F. Batista, L. Coheur, C. Parra, … M. Forcada, 365–374. European Association for Machine Translation.

Prieels, L., I. Delaere, K. Plevoets, and G. De Sutter. 2015. A corpus-based multivariate analysis of linguistic norm-adherence in audiovisual and written translation. *Across Languages and Cultures* 16 (2): 209–231. https://doi.org/10.1556/084.2015.16.2.4.

Priyatkina, A.F., E.A. Starodumova, G.N. Sergeeva, et al. (eds.). 2001. *A Russian dictionary of functional words [Slovar' sluzhebnyh slov russkogo jazyka]*. Vladivostok: Far-East State University Press.

Puurtinen, T. 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. https://doi.org/10.1093/llc/18.4.389.

Rabadán, R., B. Labrador, and N. Ramón. 2009. Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment. *Babel* 55 (4): 303–328. https://doi.org/10.1075/babel.55.4.01rab.

Rabinovich, E., and S. Wintner. 2013. Unsupervised identification of tr association for computational linguistics anslationese. *Transactions of the Association for Computational Linguistics* 3: 419–432. https://doi.org/10.1162/tacl_a_00148.

Redelinghuys, K. 2016. Levelling-out and register variation in the translations of experienced and inexperienced translators: A corpus-based study. *Stellenbosch Papers in Linguistics* 45: 189–220. https://doi.org/10.5774/45-0-198.

Santini, M., A. Mehler, and S. Sharoff. 2010. Riding the rough waves of genre on the web concepts and research questions. In *Genres on the web: Computational models and empirical studies*, vol. 42, ed. A. Mehler, S. Sharoff, and M. Santini, 3–30. Springer Science & Business Media.

Santos, D. 1995. On grammatical translationese. In *Proceedings of the 10th Nordic conference of computational linguistics (NODALIDA 1995),* ed. K. Koskenniemi, 59–66. University of Helsinki.

Sharoff, S. 2018. Functional text dimensions for annotation of web corpora. *Corpora* 13 (1): 65–95. https://doi.org/10.3366/cor.2018.0136.

Shvedova, N. (ed.). 1980. *Russian grammar*. Moscow, Science [Nauka].

Sominsky, I., and S. Wintner. 2019. Automatic detection of translation direction. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2019)*, ed. R. Mitkov and G. Angelova, 1131–1140. INCOMA Ltd. https://doi.org/10.26615/978-954-452-056-4_130.

Specia, L., G.H. Paetzold, and C. Scarton. 2015. Multi-level translation quality prediction with QUEST++. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, ed. H. Chen, and K. Markert, 115–120. Association for Computational Linguistics. https://doi.org/10.3115/v1/p15-4020.

Straka, M., and Straková, J. 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,* ed. D. Zeman, J. Hajic, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, 88–99. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-300.

Stymne, S. 2017. The effect of translationese on tuning for statistical machine translation. In *Proceedings of the 21st Nordic conference of computational linguistics,* ed. J. Tiederman, 241–246. Linköping University Electronic Press.

Teich, E. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts.* (TTCP, 5). Berlin: De Gruyter Mouton.

Toury, G. 1995. *Descriptive trantslation studies-and beyond*. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.4.

Vela, M., and E. Lapshinova-Koltunski. 2015. Register-based machine translation evaluation with text classification techniques. In *Proceedings of the 15th machine translation summit (Vol. 1: MT Researchers' Track),* ed. Y. Al-Onaizan, and W. Lewis, 215–228. Association for Machine Translation in the Americas.

Volansky, V., N. Ordan, and S. Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30 (1): 98–118. https://doi.org/10.1093/llc/fqt031.

Xiao, R., L. He, and Y. Ming. 2010. In pursuit of the third code: Using the ZJU corpus of translational Chinese in translation studies. In *Using corpora in contrastive and translation studies*, ed. R. Xiao, 182–214. New Castle: Cambridge Scholars Publishing.

Zanettin, F. 2013. Corpus methods for descriptive translation studies. *Procedia-Social and Behavioral Sciences* 95: 20–32. https://doi.org/10.1016/j.sbspro.2013.10.618.

Zhang, M., and A. Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the fourth conference on machine translation (Volume 1: Research Papers),* ed. O. Bojar, R. Chatterjee, Ch. Federmann, M. Fishel, Y. Graham, ... K. Verspoor, 73–81. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5208.

Maria Kunilovskaya PhD in contrastive linguistics, has an extensive experience in translator education and corpus linguistics. Her research is recently focused on human translation quality estimation, translationese studies and varieties, building and exploiting comparable and parallel corpora. Maria's other interests include computational and empirical approaches to comparing languages and understanding translation processes.

Gloria Corpas Pastor PhD in English philology, Professor in translation, interpreting and translation technology and an active member of several international and national editorial and scientific committees. She is the head of the LEXYTRAD research group. Her research interests include specialised translation and interpreting technologies, corpus-based translation studies, phraseology, lexicography and terminology.