# Computational Approaches to Register as a Factor in English-to-Spanish Translation

**Kateryna Poltorak**
University of Malaga
Malaga, Spain
katrinpoltorak@uma.es

**Maria Kunilovskaya**
University of Wolverhampton
Wolverhampton, UK
maria.kunilovskaya@wlv.ac.uk

## Abstract

Translations are known to be different from originally-authored texts in the target language in their lexical, structural, and stylistic patterning. This project aims to explore the impact of register on the properties of translations. Do some translated registers have a higher propensity to translationese than others? Are there translationese indicators that are shared across all registers or specific for individual registers? What is the contribution of the opposite tendencies of shining-through and over-normalisation in the overall translationese shift in each register? The answers to these questions can shed light on register-related specificity of translations in Spanish-English language pair.

## 1 Introduction

The grammar and lexis vary from language to language, and this is inevitably reflected in the product of translation. Cross-lingual asymmetry and other factors in the translation process, such as cognitive load and social norms, have an impact on translated language and give rise to its peculiar properties commonly referred to as translationese. A lot of effort was invested into the study of translationese, including approaches ranging from manual inspection of selected features to enhanced machine learning methods. Translationese is revealed through the comparison of translations and the target language (TL) norm. Yet, to understand why translations deviate from the expected norm, it is necessary to include their source texts in the equation. It allows to explain linguistic specificity of translations by either gravitation towards the TL norm or by interference from the source language. In this work, we put sources, translations and comparable non-translation in the TL into the same feature space to explore the impact of register on the properties of English-into-Spanish translations. In what follows, we first summarise previous translationese studies (Section 2). Then, Section 3 provides details on our data and methodology, including data collection, feature extraction and machine learning setup. Exploratory visualisations of the data and outcomes of translationese and register classifications are reported and discussed in Section 4. Finally, we summarise our findings and draw conclusions.

## 2 Theoretical background

The properties of translations that make them deviate from the expected TL norm are collectively known as translationese (Gellerstam, 1996). Ever since Baroni and Bernardini (2006) demonstrated that a text classification algorithm, Support Vector Machines (SVM), could distinguish translations from non-translations more effectively than human experts, translationese studies used machine learning (ML) as a research method. The same supervised ML approach was employed by Ilisei et al. (2010), who found evidence for one particular manifestation of translationese: translated texts in medical and technical domains were simpler in structure than comparative non-translations. Subsequent research established that the properties of translations were shaped by a number of factors, first and foremost, by the source language (SL) Koppel and Ordan (2011); Volansky et al. (2015); Rabinovich et al. (2017), but also by register (Kunilovskaya and Corpas Pastor, 2021), competence level (De Sutter et al., 2017), method of translation (Lapshinova-Koltunski, 2017, who explores differences between human and machine translations).

Volansky et al. (2015) used text classification methods to compute the differences between original texts in English and translations from 10 European languages into English. The scholars elaborated a set of 32 features which they grouped around four translationese hypotheses - explicitation, simplification, normalisation and interference.



*New Trends in Translation and Technology*, pages 82–90
4–6 July, 2022. 

They convincingly demonstrated that the features associated with *interference* were the most prominent translationese indicators.

Aware of the controversies associated with assigning linguistic features to specific translationese trends in a top-down matter, Kunilovskaya and Corpas Pastor (2021) preferred to explore translationese effects in a bottom-up way by comparing frequency patterns in translations, their sources and non-translations (i.e. originally-written texts from the same TL register. For their study of register variation in English-to-Russian translation, the authors designed a set of morphological, syntactic, and text-level features shared by both languages. Roughly following their methodology, we aim to extend it to the comparison of two registers in another language pair. To the best of our knowledge, cross-register properties of English-to-Spanish translations have not been an object of special investigation. The novelty of this study is in putting the performance of traditional hand-engineered features into the perspective of the results on pre-trained contextualised sentence embeddings. We expect theoretically-motivated explicit features to return competitive classification results while having unparalleled interpretative potential. We demonstrate that feature selection and feature analysis can shed light on the register specificity of out-of-English Spanish translations.

## 3 Methodology

### 3.1 Data

Our textual data comes from two English-to-Spanish parallel corpora (and comparable non-translations in Spanish) that contain fiction and parliamentary speeches. Debates corpus was sourced from the EuroParl-UdS, a collection of parliamentary speeches specifically adapted for translationese studies (Karakanta et al., 2018). It includes a comparable subset of non-translations in Spanish indicated as 'ref' (reference) in Table 1.

To get a fiction subcorpus, we extracted seven novels from a collection of out-of-copyright novels in English and their translations into Spanish[1]. The reference corpus for fiction includes ten novels extracted from the source language side of Spanish-to-Russian parallel subcorpus included in the Russian National corpus[2]. To obtain a reasonable number of observations for ML experiments, each novel was partitioned chapter-wise to get the corpus parameters that appear in Table 1.

| Subcorpus | Words | Sentences | Texts |
|---|---|---|---|
| **Debates** | | | |
| src | 3,844,826 | 152,638 | 540 |
| tgt | 3,914,618 | 144,853 | 540 |
| ref | 6,335,046 | 189,526 | 1031 |
| **Fiction** | | | |
| src | 654,242 | 30,094 | 144 |
| tgt | 587,734 | 33,365 | 144 |
| ref | 944,489 | 53,198 | 258 |

Table 1: Details on the corpus
(src = source, tgt = target, ref = non-translations in TL)

### 3.2 Features Extraction and Alternative representations

This study aims to compare the properties of translations between two registers by looking at the linguistic features that were most useful in the automatic classification of labelled documents. To obtain a linguistically interpretable representation of the textual data, we developed a set of 51 morphological, syntactic and textual features that can be extracted from UD annotations[3] of the documents. Each document received a multidimensional vector, where each component had a value for a particular linguistic feature (typically, normalised frequencies or ratios). The features varied in terms of extraction procedures. Values for most of them were obtained by matching the morphological and syntactic tags, and their combinations from UD annotations. This subset was complemented by common and attested textual parameters, established as effective translationese indicators in previous research, such as type-to-token ratio (TTR) and sentence length.

For feature extraction in this part, we adapted the rules developed for English-Russian language pair and made available in Kunilovskaya et al. (2021) [4]. Generally, we proceeded from the assumption that the bigger the pool of available features, the higher the chances to discover linguistically- and translationally-interesting translationese indicators.

Along with the features based on morphological and syntactic tagging, we used the frequencies of discourse markers from predefined search lists as proposed in Kunilovskaya and Corpas Pastor

[1] https://farkastranslations.com/bilingual_books.php
[2] https://ruscorpora.ru/en/
[3] https://universaldependencies.org/introduction.html
[4] https://github.com/kunilovskaya/translationese45

(2021). They were classified into additive, adversative, causative, temporal sequential and epistemic. When compiling the lists of markers for Spanish, we relied on textual corpus query systems, translation textbooks and other linguistic resources. The list of the connectives was then checked against the respective grammar reference books for both languages. We included single-word and multi-word discourse markers and extended lexical and structural representation of particular items, allowing for some degree of variation (e.g. *in my view* and *in our view* for markers of epistemic stance; *second* and *secondly* for sequential markers). For example, the group of adversative connectives for English included 46 items such as *however*, *in spite of*, etc; the comparable Spanish list had items like *a diferencia*, *al contrario*, *ahora bien*, etc. Orthography and punctuation were also used to distinguish between different patterns of discourse connectives (e.g. *furthermore* and *Furthermore* were matched separately).

The majority of the features were normalised to the number of sentences, including discourse markers, verb forms, simple sentences, and number of clauses per phrase. Some features have their own normalisation basis. For example, nouns in the functions of subject, object, or indirect object were normalised to the total number of the occurrences of each of these types in the text Kunilovskaya and Corpas Pastor (2021).

To offer a challenging competition for the morphosyntactic representation, detailed above, we generated two automatic representations. First, we used a *tf-idf* vectoriser as implemented in *scikit-learn* [5], a well-known library for ML in Python. Each text was represented as a vector of tfidf-weighted frequencies of 5000 most frequent character 5-grams seen in the whole collection. All other settings were set to the library defaults.

Second, each document in the corpus was represented by a vector generated by averaging sentence vectors, obtained from a pre-trained contextualised embedding model from *Hugging Face* repository, *stsb-xlm-r-multilingual* [6]. STSB-XLM-R-M is a multilingual XLM-R model fine-tuned on semantic textual similarity task using Sentence-BERT (SBERT), a state-of-the-art framework to compute sentence representation (Reimers and Gurevych, 2019).

[5]https://scikit-learn.org/stable/

[6]https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual

### 3.3 Experimental Setup

Our experiments involve a set of binary text categorisation tasks to automatically distinguish translations and non-translation in each register, and non-translated registers against each other in each language. To get insights into the importance of individual features for these distinctions, we relied on feature selection and univariate statistic analysis. SVM was shown to be superior to other ML methods on this task (Ilisei et al., 2010; Van Halteren, 2008). In our study, an SVM classifier is set with a linear kernel and default $C$=1.0. All input data was scaled with *scikit-learn* Standard Scaler before experiments.

In some of our experiments, classes were not well-balanced: for example, in the register classification in non-translated Spanish, we had 1027 instances for debates and only 248 instances for fiction. To counteract a possible classification bias, we set *class_weight* option to 'balanced' in all experiments.

The results of classification were evaluated in *7-fold cross-validation* setting. We partitioned each novel in the fiction subcorpus into multiple chunks. This corpus structure warrants the implementation of "novel-aware" folds to avoid training and testing on the chunks from the same novel. To achieve this, we used *Stratified GroupKFold* strategy from the scikit-learn library. In each data split, the model is trained on chunks from six books, while the test set contains the chunks from the seventh book that was unseen in training.

The same approach to classifier evaluation (*Stratified GroupKFold*) was used to obtain results on *tf-idf* representations and on text vectors generated by mean-pooling of sentence embeddings from a *STSB-XLM-R-M* model.

To make sure that SVM returns a good performance on a dataset where the ratio between the number of observations and the number of features is relatively small (i.e. for high-dimensional representations), we designed a simple one-layer neural model which uses *Adam optimiser* and *binary cross-entropy* as the loss function. *EarlyStopping class* was set up to monitor accuracy at the validation stage. The neural model was compiled with one dense hidden layer made of 16 neurons. This layer was activated with *ReLU (Rectified Linear Units) function*, a piece-wise linear function that outputs zero if it receives a negative value and returns the input directly if it is positive. The output

layer had one output node: it generated the probability of class 1, i.e. positive class.

## 4 Results and Discussion

### 4.1 Visualisations of the Data

To obtain a general understanding of the locations of translated and non-translated registers in the vector space created by the three numeric representations (morphosyntactic features, tf-idf, sentence embeddings), Principal component analysis (PCA) was used to project the data into a two-dimensional space. Using PCA-reduced 2D vectors, we generated scatter plots for all data and for each register separately. Naturally, a 2D projection introduced distortions to the actual similarity relations between data points but it allowed us to get a general view of how well the features capture the categories in the data.

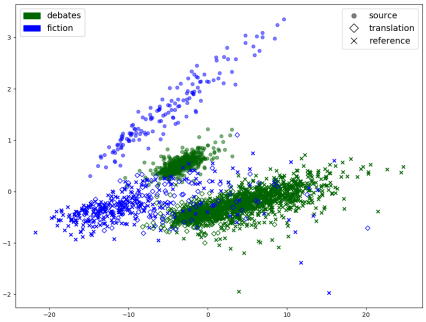


Figure 1: PCA projection for UD features

Figure 1 represents all six categories of texts: two registers and three text types. Each marker in the graph is a text located by two components returned by PCA of 51-dimensional UD-based vectors (morphosyntactic features). The overwhelming amount of variance in the data is accumulated in PCA Dimension 1 (98%). The values for texts on this dimension are represented by the horizontal axis (x-axis). Using 51 morphosyntactic features, it is difficult to say which aspect of the dataset is captured in this first PCA component: translational status or register.

On the one hand, translations indicated by empty diamonds seem to be shifted horizontally from comparable non-translations (crosses) in each register, even though in different directions: for fiction (blue) to the right and for debates (green) to the left. Translations seem to slightly gravitate toward the center of the x-axis continuum. This can be indicative of levelling-out of register distinctions in translation: if so, translated registers should be more difficult to tear apart by an automatic classifier and would yield lower classification results than non-translations.

On the other hand, the areas occupied by the two registers, at least in the TL, seem to be more clearly defined on x-axis (see the lower two clouds of blue and green markers). At the same time the source language registers, represented by solid dots, are not that well separated horizontally, with source in fiction demonstrating a lot of unfocused variability. The y-axis values can be said to accumulate language contrast: sources (solid dots in either colour) are mostly located in the upper part of the graph, and translations and comparable reference texts in the TL are in the lower part of the graph.

By way of comparison, we can look at how the six categories are located in the 2D PCA projections on the alternative vectorisations (tf-idf vectors and mean-pooled sentence vectors generated by a pre-trained model).

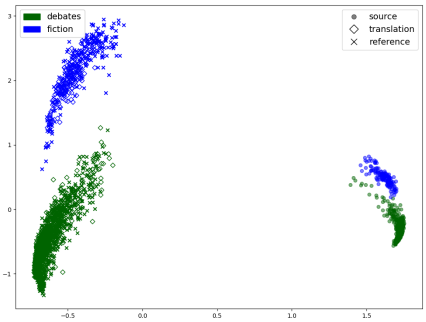


Figure 2: PCA projection of tf-idf vectors

In Figure 2, based on PCA 2D reduction of tf-idf vectors, we see four distinct compact areas of markers in the form of a comma in two colours corresponding to the two registers. If we consider the x-axis, all documents in the TL are on the left, and documents in the SL are on the right. Notably, the distance between languages is greater than the distance between registers reflected at the y-axis. This shows that tf-idf captured the language and register distinction quite well, while the text type distinction focused in this work - translations vs non-translations - is amiss. Note that tf-idf representations were trained separately for each language. It is not surprising that language contrast overshadows all other distinctions. Actually, tf-idf

comparisons make sense only for documents in the same language.

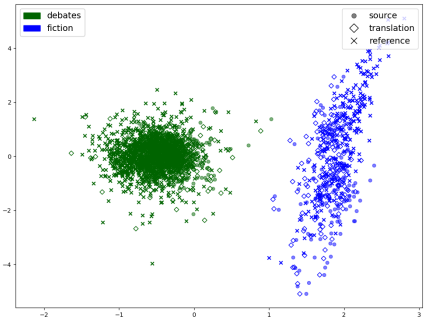


Figure 3: PCA projection of stsb_xlm_r vectors

Figure 3 has a 2D projection of the STSB-XLM-R-M-based embeddings for our documents. It is a cross-lingual model, and, unlike tf-idf, it puts documents in different languages into the same vector space. The x-axis shows the difference in registers: a circle-shaped debates cloud and an ellipsis-shaped blue cloud of fiction texts can be clearly seen. There is no clear distinction between the three translationese-related types of texts on either the x-axis or the y-axis. From this, we can deduce that the PCA reduction of contextualised embeddings oversimplifies the text distinctions that might be reflected in these vectors and captures only the strongest parameter - register.

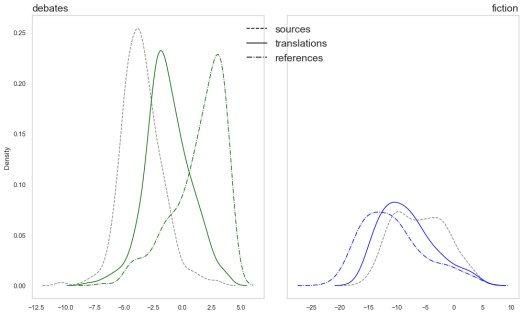


Figure 4: Distribution of values on PCA dimension 1 (morphosyntactic features)

Figure 4 provides a different perspective on the PCA analysis of our corpus in its morphosyntactic representation. It shows the density of the values on the first dimension of the PCA-reduced data. The x-axis has a continuum of values, while the y-axis reflects the frequency of these values as observed in the dataset. In fact, it is a smoothed version of a histogram used to show the distribution of a variable. The plots in both panels (debates on the left, fiction on the right) clearly capture the distinction between the three text types. Importantly, it can be seen that translations are located between their sources and comparable non-translation in both registers, confirming their 'third code' nature (Frawley, 1984).

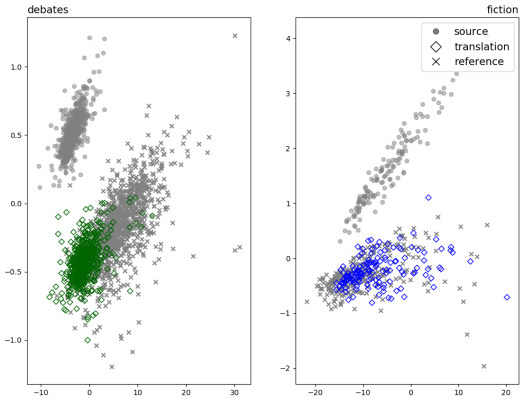


Figure 5: Sources, translations and non-translations by register (UD features)

To get a clearer view of the vector-space locations of the three text types in each register, we produced a separate scatter plot for each register (see Figure 5). These zoomed-in representations demonstrate that sources are well-separated from texts in the TL along the y-axis in both registers. Translations in the left-hand debates panel (shown in green) seem to form a more isolated and compact cloud shifted away from the area taken by non-translations in the TL (grey crosses) than is the case for translated fiction. Based on this observation, we can expect higher translationese classification results for debates than for fiction.

### 4.2 Register Classification by Language

Before our main experiment with translationese classification, we tested whether the 51 morphosyntactic features could detect intra-linguistic register contrast based on non-translated texts in English (sources) and non-translated texts in Spanish (references) (see Table 2). It is interesting to compare the outcomes of on sources and references to register distinctions in translated Spanish (Table 3). The results of three binary classifications are reported for a linear SVM classifier with default settings, and for morphosyntactic features as input only.

The results for the experiments to distinguish non-translated registers are systematised in Table 2.

Registers in English source texts were classified with 100% accuracy. Debates and fiction originally

| | acc (%) | F1 (%) |
|---|---|---|
| English | 100 | 100 |
| Spanish | 98.9 | 98.2 |

Table 2: Accuracy and F1-score for register classifications on non-translated documents

| | acc (%) | F1 (%) |
|---|---|---|
| translated Spanish | 99.6 | 99.4 |

Table 3: Accuracy and F1-score for register classification in translated Spanish

written in Spanish were predicted with an accuracy of 98.9%. In the latter experiment, there were 13 misclassified items in total out of 1275 observations: four documents from debates were predicted as fiction and 10 instances of fiction were predicted as debates. Quite surprisingly, translated Spanish registers were even more distinguishable than non-translated ones, achieving an accuracy of 99.6% with 3 misclassified instances out of 680 observations in this experiment. These results indicate that the morphosyntactic feature captured the register distinctions very well in both languages, but did not support levelling-out hypothesis.

### 4.3 Translationese Classification

The purpose of the second group of experiments was to see whether our feature set worked well for capturing translationese across registers, which was key to our study.

To render our results comparable across various representations, we report accuracy and F1-score for the SVM and the neural classifier on each one of them. Tables 4, 5, 6 and 7 present the results for two classifiers and three types of vectors in two registers.

| | acc (%) | F1 (%) |
|---|---|---|
| msynt | 96.6 | 96.2 |
| tf-idf | 99.5 | 99.4 |
| embeddings | 98.4 | 98.2 |

Table 4: Debates: Translationese classification by a linear SVM

On **morphosyntactic features** in the debates subcorpus, the linear SVM classifier achieved 96.6% accuracy with 96.2% F1-score while in the fiction subcorpus the same settings gave lower results of 81.2% accuracy and 77.3% F1-score. The neural classifier returned very similar results (debates: 96.6% accuracy with 96.2% F1-score; fiction: 81.1% accuracy and 78.2% F1-score). It is an expected outcome because the input to both classifiers was exactly the same, and we did not perform any fine-tuning in the neural network setup.

| | acc (%) | F1 (%) |
|---|---|---|
| msynt | 96.6 | 96.2 |
| tf-idf | 99.2 | 99.1 |
| embeddings | 97.3 | 97.0 |

Table 5: Debates: Translationese classification by a neural model

If we consider the **tf-idf representation**, both classifiers detected translationese with almost perfect precision in the debates subcorpus. The linear classifier achieved 99.5% with F1-score of 99.4% and the numbers for the neural classifier were 96.4 % of accuracy with F1-score 93.9%. The outcome for literary texts was slightly lower with the linear classifier (96.4% of accuracy with F1-score of 93.9%) and noticeably lower for the neural network (88.8% accuracy and 86.3% F1-score). This indicates that 2D PCA projections of 5K features are but a crude oversimplification of this representation; the information about translational status of a text is distributed across many features that cannot be adequately rendered by just two factors (two PCA dimensions).

| | acc (%) | F1 (%) |
|---|---|---|
| msynt | 81.2 | 77.3 |
| tf-idf | 96.4 | 93.9 |
| embeddings | 96.6 | 96.0 |

Table 6: Fiction: Translationese classification by a linear SVM

| | acc (%) | F1 (%) |
|---|---|---|
| msynt | 81.1 | 78.2 |
| tf-idf | 88.8 | 86.3 |
| embeddings | 94.1 | 91.6 |

Table 7: Fiction: Translationese classification by a neural model

**Contextualised word embeddings** also showed high results on the texts from the debates subcorpus. The SVM classifier spotted translated texts

with 98.4% of accuracy and 98.2% of F-score and the neural classifier returned 97.3% accuracy with the F-score of 97.0%. The results on the fiction subcorpus are also very convincing: 96.6% of accuracy and 96.0% of F-score showed by the linear SVM classifier and 94.1% with F-score of 91.6% achieved by the neural classifier.

The results of all the experiments suggest that three representations with both classifiers achieved high accuracy on the texts from the debates subcorpus, and slightly lower accuracy on the fiction subcorpus. It is in line with the intuition: fiction in translation is expected to be more adapted to the TL norm and to exhibit less influence of the source text and language due to less emphasis on accuracy and higher aesthetic standards in this register.

### 4.4 Best Translationese Indicators by Register

The goal of this experiment was to identify the features that are most associated with either translated or non-translated text categories using *analysis of variance* (ANOVA) and to compare them across registers. Our first observation confirms findings by Kunilovskaya and Corpas Pastor (2021) for English-Russian translated registers: each register generated its own type of translationese, i.e. there were no universal translationese indicators. The results for the top five translationese indicators were worse for fiction (9% down from the full feature set) than for debates (6% down from the full feature set). This finding suggested that translationese properties were less expressed in fiction, i.e. translations were more statistically in line with the expected TL register norm.

### 4.5 Translationese Effects based on Univariate Analysis

We also ran a series of univariate analyses on each UD feature to see whether a particular linguistic feature could reveal shining-through effect, normalisation, or cannot be considered a translationese feature at all, given our definition of translationese. First of all, we extracted the averaged normalised frequencies of each feature in: (i) original texts (English); (ii) translations into Spanish; (iii) reference texts in the TL (Spanish), for each morphosyntactic feature, for each register subcorpora separately.

Then, we ran a significance test (t-test) on the pairs of corpora: English sources vs Spanish non-translations (language gap); Spanish translations vs Spanish non-translations (translationese) and English sources vs Spanish translations (to establish SL/TL translationese). For the features that had significant distinctions between the corpora, we calculated Cohen's D coefficient to obtain the effect size of the differences. It is interpreted as follows. If Cohen's D is closer to one, the differences are considerable; a score of 0.5–0.6 is high enough, and a score of 0.2 and less indicates that, although the differences in frequencies are significant according to t-test, it is mostly due to the available amount of data, not because the differences are big.

An important criterion that we took into account was the language gap. Even if the algorithm did not detect any translationese, if there was language contrast for it, a particular feature could be considered fully adapted to the TL norm. If neither translationese nor language gap was detected, a particular feature was considered useless. Following this logic, we grouped the results of the univariate analysis for the fiction and debates corpora according to the features that pointed to (1) *shining-through* and (2) *normalisation* in translation, and also listed (3) *language independent* features.

*The number of nouns per sentence (nn)* was one of the features that revealed shining-through effect in the fiction subcorpus and thus underlined the nominal nature of the Spanish language which was emphasised by previous contrastive studies that mostly investigated mass media discourse (Casado-Velarde, 1978; Casasús Josep and Núñez Ladevéze, 1991; Ladevéze, 1993; Núñez Ladevéze, 2011; Ruiz, 2010), scientific and technical discourse (Albentosa Hernández, 1997). Interestingly, in the debates, this feature indicated normalisation in translation. Among the whole feature set, 33 features were associated with shining-through in translations of parliamentary speeches and only 13 features contributed to this effect in the fiction subcorpus. Based on our results, we can conclude that translators of the parliamentary texts were more concerned about preserving the original structures, while sacrificing the norm of the TL. On the contrary, translators of fiction had more freedom in terms of language resources so they would not just convey the meaning, but adapt it to the target reader in a natural way, thus reducing the shining-through effect in translation.

Some of the features indicated a trend to normalisation in translation, i.e. in cases when translations from English gravitated towards the norm of the Spanish language. Interestingly, finite verbs worked for both subcorpora: in both registers there

was a difference between frequencies of finite verbs in the originals and the reference texts, which pointed to the language gap, while the number of finites in translations and the references were almost the same. This is a case of full adaptation to the TL norm.

Normalisation in translations of the parliamentary texts was seen on the number of nouns per sentence (*nn*). An illustration of normalisation in translation could be the English verb *legislate* that was rendered in Spanish as *introducir legislación*. Yet, it is difficult to give convincing examples of translationese effects because translationese is a probabilistic phenomenon which is spread in the text. In professional translations, it does not lead to obvious deviations from the expected TL norm. Most probably what our classifiers are catching are cases of positive transfer, i.e. overreliance on the structures that exist in both languages to the detriment of Spanish-specific expressions.

We also detected features that captured SL/TL independent translationese. These were the cases when there was no language gap but there was a significant difference in the averaged normalised frequencies between (i) translations and source and (ii) translations and reference texts. For example, it was expected that Spanish translations would contain more negative particles, because in Spanish, unlike English, double negation occurs. Yet, it was surprising that the number of negative particles in translations was even higher than in the comparable texts of the same register in Spanish.

## 5 Conclusions

In this study, we investigated the effect of registers on the properties of translations from English into Spanish. Our first experiment established that the proposed feature set was effective in capturing the intra-linguistic register contrast in both English and Spanish. The classifiers showed an accuracy of 98.9% for debates and fiction texts originally written in Spanish while the registers in the English language were predicted with 100% accuracy. Our main focus was on the second series of experiments, where we wanted to see whether our feature set worked well for capturing translationese across registers. On the morphosyntactic features, the linear SVM classifier performed well on the debates subcorpus while the same settings gave lower results on the fiction subcorpus. The neural classifiers showed almost the same dynamics: they were used as a sanity check option to confirm the results from SVM. The results on alternative vectorisations were used to confirm the trends revealed by the morphosyntactic representation: there was more translationese in debates than in fiction. We also conducted a series of univariate analyses on the whole feature set to trace linguistic indicators that revealed shining-through, normalisation, or SL/TL-independent translationese trends. Our results on the English-Spanish pair showed that 14 indicators were associated with the shining-through effect in the fiction subcorpus (e.g. sentence length, the number of nouns per sentence, mean dependency distance, etc.), and 34 features pointed to the same phenomenon in the debates subcorpus (e.g. sentence length, word length, mean hierarchical distance). This confirms the findings in Kunilovskaya and Corpas Pastor (2021) who proved that translated literary texts in the English-to-Russian language pair were less prone to the shining-through effect.

## References


José Ignacio Albentosa Hernández. 1997. La sustantivación en el discurso científico en lengua inglesa. *Cauce, 1997-1998,(20-21): 329-344*.

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

Manuel Casado-Velarde. 1978. La transformación nominal, un rasgo de estilo de la lengua periodística.

Maria Casasús Josep and Luis Núñez Ladevéze. 1991. Evolución y análisis de los géneros periodísticos" en josep maria casasús y luis núñez ladevéze. *Estilos y géneros periodísticos*.

Gert De Sutter, Bert Cappelle, Orphée De Clercq, Rudy Loock, and Koen Plevoets. 2017. Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translations. *Linguistica Antverpiensia*, 16:25–39.

William Frawley. 1984. Prolegomenon to a theory of translation. *Translation: Literary, linguistic and philosophical perspectives*, 159:175.

Martin Gellerstam. 1996. Translations as a source for cross-linguistic studies. *Lund studies in English*, 88:53–62.

Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *International Conference*

*on Intelligent Text Processing and Computational Linguistics*, pages 503–511. Springer.

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. Europarl-uds: Preserving and extending metadata in parliamentary debates. *ParlaCLARIN: Creating and Using Parliamentary Corpora*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326.

Maria Kunilovskaya and Gloria Corpas Pastor. 2021. *Translationese and register variation in English-to-Russian professional translation*, pages 133–180. Springer Nature Singapore Pte Ltd.

Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski, and Ruslan Mitkov. 2021. Fiction in russian translation: A translationese study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 734–743.

Luis Núñez Ladevéze. 1993. *Teoría y práctica de la construcción del texto: investigación sobre gramaticalidad, coherencia y transparencia de la elocución*. Editorial Ariel.

Ekaterina Lapshinova-Koltunski. 2017. Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. In Gert De Sutter, Marie-Aude Lefer, and Delaere Isabelle, editors, *Empirical Translation Studies. New Theoretical and Methodological Traditions*, Trends in Linguistics. Studies and Monographs, vol. 300, pages 207–234. De Gruyter Mouton, Berlin/Boston.

Luis Núñez Ladevéze. 2011. *Métodos de redacción periodística y fundamentos del estilo*. Madrid: Síntesis, DL 1993.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ramón González Ruiz. 2010. Gramática y discurso: nominalización y construcción discursiva en las noticias periodísticas. In *Estrategias argumentativas en el discurso periodístico*, pages 119–146. Peter Lang.

Hans Van Halteren. 2008. Source language markers in europarl translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 937–944.

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