EUROPEAN MASTERS IN TECHNOLOGY FOR TRANSLATION AND INTERPRETING

Module: MA Dissertation

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

Enfoques Computacionales del Registro como Factor en la Traducción del Inglés al Español

Student: Kateryna Poltorak

Supervisor 1: Dr Maria Kunilovskaya

Supervisor 2: Dr María Rosario Bautista Zambrana

Contents

Declaration

Declaration/Declaración

EUROPEAN MASTERS IN TECHNOLOGY FOR TRANSLATION AND INTERPRETING (EM TTI)

University of Wolverhampton

University of Malaga

Name/nombre: Kateryna Poltorak

Title/Título: Computational Approaches to Register as a Factor in English-to-Spanish Translation

Module Name/Nombre de la asignatura: MA Dissertation

Supervisors/Directores:

Dr Maria Kunilovskaya

Dr Maria Rosario Bautista Zambrana

Declaration:

(i) This work or any part thereof has not previously been presented in any form to the University or to any other institutional body whether for assessment or for other purposes. Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person. (ii) It is acknowledged that the author of any project work shall own the copyright. However, by submitting such copyright work for assessment, the author grants to the University a perpetual royalty-free licence to do all or any of those things referred to in section 16(i) of the Copyright Designs and Patents Act 1988 (UK) (viz: to copy work; to issue copies to the public; to perform or show or play the work in public; to broadcast the work or to make adaptation of the work) and Ley de propiedad Intelectual (Ley 2/2019, de 1 de marzo) (Spain). (iii) This project did not involve direct contact with human subjects, and hence did not require approval from the Faculty Ethics Committee.

Declaración:

(i) Este trabajo o cualquier parte del mismo no ha sido presentado previamente en ninguna forma a la Universidad o a cualquier otro organismo institucional ya sea para su evaluación o para otros fines. Salvo reconocimientos expresos, referencias y/o bibliografías citadas en el trabajo, confirmo que el contenido intelectual del mismo es fruto de mi propio esfuerzo y de ninguna otra persona.

(ii) Se reconoce que el autor de cualquier trabajo del proyecto es el propietario de los derechos de autor. Sin embargo, al presentar dicho trabajo con derechos de autor para su evaluación, el autor concede a la Universidad una licencia perpetua y libre de derechos para realizar todas o cualquiera de las cosas mencionadas en la sección 16(i) de la Ley de Derechos de Autor, Diseños y Patentes de 1988 (Reino Unido) (a saber: copiar el trabajo; emitir copias al público; representar o mostrar el trabajo en público; emitir el trabajo o

hacer una adaptación del mismo) y Ley de propiedad Intelectual (Ley 2/2019, de 1 de marzo). (iii) Este proyecto no implicó el contacto directo con seres humanos, por lo que no requirió la aprobación del Comité de Ética de la Facultad.

Date/ Fecha: 18-05-22

Students signature/ Firma del estudiante:

Acknowledgements

My BIGGEST gratitude goes to my first supervisor Dr Maria KuniLOVEskaya, who supported me throughout the whole EMTTI journey, who was always there for me (including on holidays and weekends) if I had doubts and needed advice, and whose kind guidance and expertise helped me to do this research. I would like to acknowledge the help of my second supervisor Dr María Rosario Bautista Zambrana, her kindness and expert advice. I am deeply grateful to the coordinators and the Consortium of the EMTTI programme for putting so much effort in making it all happen. I am enormously grateful to Prof. Mitkov, Prof. Corpas, and all the academic staff at the University of Wolverhampton and the University of Malaga. And last but not least, a massive thanks goes to my fellow students who supported me during these two years.

Abstract

Translations are known to be different from the originally-authored texts in the target language in their lexical, structural and stylistic patterning. The properties of translations that make them deviate from the expected target language norm are collectively known as translationese. Koppel and Ordan (2011) spoke about the "dialects" of translationese by which they referred to translations from different source languages into the same target language. This project adopts the quantitative and descriptive definition of translationese which is manifested in several assumptions underpinning our methodology. (1) The higher the quality of the automatic classification which distinguishes translations and comparable non-translations, the more translationese is carried by the translations. (2) A linguistic feature is recognised as a translationese indicator if its frequency of occurrence in translations is significantly different from what is observed in comparable originally-authored texts in the target language (non-translations) in the target language. (3) Explanatory analysis of translationese needs to factor in feature frequencies in the aligned source texts, extending the corpora comparison to a three-partitive relation: sources, targets and non-translations.

This project aims to explore the impact of register on the properties of translations. Do some translated registers have 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 a given language pair. Our research design avoids averaging across language variations associated with register and accounts for the linguistic distance between the source and expected target norm in interpreting the findings. It helps in capturing deviations that cannot be explained by the pull from either source or target languages but can be attributed to socio-cultural/professional norms operating in individual registers or cognitive (non-linguistic) factors that cut across all registers. Our findings are relevant for alerting humans and machines to the deviations from the expected target language norm typical for translators and observed in the training data.

Preliminary results in the cross-validation setting indicated that based on the full 51-dimensional feature vectors, translations of fictional books were less distinct from non-translations (accuracy 78% with a macro F1-score of 0.60). It means that they exhibited less translationese than for example translations of parliamentary speeches (96% with a macro F1-score of 0.89), which is in line with the expected. In our experiments, the classes are not strictly balanced, so F-score gives a better evaluation of the classifiers performance. Note, that for the second experiment the results were only slightly above the chance level of 61% (or F1-score of 0.56). In the SVM setting, the classification results using the top five features selected according to analysis of variance (ANOVA)

test, are only a few per cent lower than those on the full feature vector, and they are not shared across registers. For example, the features with the most predictive frequency differences in fiction are sentence length, coordinating conjunctions, mean dependency distance, attributive phrases and type-to-token ratio based on content lemmas only, while in speeches this list includes frequencies of adversative connectors, possessive pronouns, verbs in the past tense, frequencies of sequential and temporal markers, and number of simple sentences.

Interestingly, this is not the case with the neural classifier which demonstrated almost perfect separation of the classes: 99.75% accuracy for fiction and 98% accuracy for speeches. These over-optimistic results should be taken with a grain of salt. We plan to have a closer look at the best translationese indicators revealed by this exploratory analysis and address the problem of insufficient textual data to represent fiction register.

Resumen

Las traducciones difieren de los textos originales en la lengua meta en su patrón léxico, estructural y estilístico. Las propiedades de las traducciones que las hacen desviarse de la norma lingüística de la lengua meta se conocen comúnmente como *translationese* o traductologías. Koppel y Ordan (2011) hablaron de los "dialectos" de *translationese*, refiriéndose a las traducciones de diferentes lenguas de origen a la misma lengua de destino. Este proyecto adopta la definición cuantitativa y descriptiva de este fenómeno que se manifiesta en varios supuestos que sustentan nuestra metodología. (1) Cuanto mayor sea la calidad de la clasificación automática que permite distinguir las traducciones de las no traducciones comparables, más *translationese* llevarán las traducciones. (1) Cuanto mayor sea la calidad de la clasificación automática que permite distinguir las traducciones de las no traducciones comparables, más traductología llevarán las traducciones. (2) Un rasgo lingüístico se reconoce como indicador de *translationese* si su frecuencia de aparición en las traducciones es significativamente diferente de la observada en textos comparables de autoría original en la lengua de destino (no traducciones). (3) El análisis explicativo de la traductología debe tener en cuenta las frecuencias de los rasgos en los textos originales alineados, ampliando la comparación de los corpus a una relación tripartita: fuentes, objetivos y no traducciones.

Este proyecto pretende explorar el impacto del registro en las propiedades de las traducciones. ¿Son algunos registros traducidos más propensos a la traductología que otros? ¿Existen indicadores de traductología comunes a todos los registros o específicos de cada uno de ellos? ¿Cuál es la contribución de las tendencias opuestas de la interferencia y sobrenormalización en el cambio general de la traductología en cada registro? Las respuestas a estas preguntas pueden arrojar luz sobre la especificidad registral de las traducciones en un par de lenguas determinado. Nuestro diseño de investigación evita hacer un promedio las variaciones lingüísticas asociadas al registro y tiene en cuenta la distancia lingüística entre la norma de origen y la norma de destino esperada a la hora de interpretar los resultados. Ayuda a captar las desviaciones que no pueden explicarse por la

tendencia de acercarse a la norma de las lengua de origen o la del destino, sino que pueden atribuirse a normas socioculturales/profesionales que operan en registros individuales o a factores cognitivos (no lingüísticos) que están presentes en todos los registros. Nuestros hallazgos son relevantes para alertar a los humanos y a las máquinas de las desviaciones de la norma esperada en la lengua de destino típicas de los traductores y observadas en los datos de entrenamiento.

Los resultados preliminares en el marco de la validación cruzada indicaron que, sobre la base de los vectores de características de 51 dimensiones, las traducciones de libros de ficción se distinguían menos de las no traducciones (precisión del 78% con una macro puntuación F1 de 0,60). Esto significa que mostraron menos traductología que, por ejemplo, las traducciones de discursos parlamentarios (96% con una macro puntuación F1 de 0,89), lo que coincide con lo esperado. En nuestros experimentos, las clases no están estrictamente equilibradas, por lo que la puntuación F ofrece una mejor evaluación del rendimiento de los clasificadores. Obsérvese que, en el segundo experimento, los resultados sólo están ligeramente por encima del nivel de azar del 61% (o una puntuación F1 de 0,56). En la configuración de la SVM, los resultados de la clasificación utilizando las cinco características principales seleccionadas según la prueba de análisis de la varianza (ANOVA), son sólo unos pocos porcentajes inferiores a los del vector de características completo, y no se comparten entre los registros. Por ejemplo, los rasgos con las mayores diferencias de frecuencia predictiva en la ficción son la longitud de la frase, las conjunciones coordinadas, la distancia media de dependencia, las frases atributivas y la relación tipo-tecla basada sólo en lemas de contenido, mientras que en los discursos esta lista incluye las frecuencias de conectores adversativos, pronombres posesivos, verbos en pasado, frecuencias de marcadores secuenciales y temporales y número de frases simples.

Curiosamente, este no es el caso del clasificador neural que demostró una separación casi perfecta de las clases: 99,75% de precisión para la ficción y 98% de precisión para los discursos. Estos resultados demasiado optimistas deben tomarse con cautela. Tenemos previsto examinar más detenidamente los mejores indicadores de traducción revelados por este análisis exploratorio y abordar el problema de la insuficiencia de datos textuales para representar el registro de ficción.

1 Introduction

It is widely acknowledged that translated texts exhibit specific characteristics as compared to originally authored texts in the target language. The grammar and lexis vary from language to language and this is reflected in the product of translation. This form of the cross-lingual asymmetry creates the residual effect on the language of translations. It is most commonly referred to as translationese. The term “translationese” was introduced by Gellerstam and is associated with the systematic influences from the source text that could be observed in the target text, or as the author pointed out – the “unmistakable

fingerprints" of the source language (Gellerstam, 1996).

A lot of effort was invested into the study of translationese, including approaches ranging from manual inspection of selected features to more technologically enhanced methods. The latest embrace corpus-based techniques that operate on large textual data and enable extracting statistical information on linguistic patterns found in translated and non-translated texts. In the early 2000th linguists started to experiment with methodologies from other disciplines. In particular, machine learning methods such as text vectorisation and text classification, started being used in translationese studies. Baroni and Bernardini (2006) were among the pioneers within translationese research community who used Support Vector Machines (SVMs) for text categorisation. In their trend-setting work, the scholars demonstrated that, based on a set of lexico-grammatical and syntactic features, an automatic algorithm can be more effective in differentiating between translations of journal articles into Italian and comparable Italian originals than humans. Subsequent research in the field embraced machine learning methods to explore the carry-over effects in translations. Ilisei (2012) applied machine learning techniques to reveal the tendencies of simplification and explicitation in texts translated into Spanish and Romanian. Bizzoni et al. (2020) analysed translationese in human and machine translation of written texts and speech transcriptions.

The established way to measure the effect of translationese is to represent the divergences between translations and non-translations through a series of features that could be automatically retrieved. Translations tend to exhibit peculiar lexical, grammatical, syntactical, and stylistic peculiarities that distinguish them from register-comparable original texts in the target language. Such linguistic manifestations of translationese can be studied computationally on large samples of data. A number of studies explore translationese features at various language levels. Rodríguez-Castro (2011) compared the frequencies of punctuation marks in articles translated from English into Spanish and original articles in Spanish. Kunilovskaya and Lapshinova-Koltunski (2019) designed an extended set of linguistically interpretable features embracing morphological, syntactic, and text-level properties to compare "translationese-prone areas" in English-to-Russian professional and student translations.

Following the research conducted by Kunilovskaya and Lapshinova-Koltunski (2019), we propose to design a comprehensive set of linguistic features to be tested on English-to-Spanish translations belonging to various registers. Cross-register translationese in the English-to-Spanish language pair has not been sufficiently investigated. The present study aims to reveal the amount of translationese manifested in Spanish translation of English texts across a number of registers. Importantly, the comparison of translations and originally-authored texts in Spanish will be thrown into perspective by contrasting the translations with their sources. This third member of the traditional two-member translationese opposition (translation vs. non-translation in the TL) is necessary to establish cross-linguistic cross-register distances between originally-authored texts in the

source and target languages. Arguably, these distances play a major role in defining the hybrid linguistic properties of translations. Besides, the reference to the source language (SL) is necessary to establish the nature of the observed translationese effect. Are the divergences from the expected TL norm motivated by the SL influence or do they come about as a result of tendencies in translation other than interference? Additionally, we intend to take the exploration of the translationese indicators further: we propose and test new translationese indicators, primarily based on morphological and syntactical features.

1.1 Aims and Objectives

The key objectives of the research are as follows:

1. Collect a multi-corpus containing parallel English-to-Spanish subcorpora and Spanish reference subcorpora for several registers.

2. Design a set of linguistically motivated morphological and syntactical features.

3. Test the features for capturing translationese in several registers to evaluate the effectiveness of the suggested features across registers.

4. Describe the translational specificity of each register, taking into account the cross-linguistic distance between the registers.

5. Interpret the results with the reference to the existing theory of translationese and throw them into perspective of the outcomes of similar studies.

1.2 Research questions

The main line of research aims at finding how different is translated Spanish across registers from comparable non-translated language.

H1: The amount of translationese is contingent on the distance between source and target language in a given register.

H2: Each register has its own translationese indicators.

Therefore, we set the following research questions:

RQ1: Do our features capture language contrast (non-translated English vs. non-translated Spanish, further referred to as source (src) and reference (ref) respectively for brevity) regardless register?

RQ2: Do our features capture intra-linguistic register contrast (e.g. English fiction vs. English debates)?

RQ3: How good are our features for capturing translationese by register?

RQ4: Which features can be considered translationese indicators for each register based on univariate analysis and which perform better in the automatic text classification setting?

2 Literature Review

2.1 Basic Concepts

2.1.1 Definitions of Translationese

As translation studies shaped into a discipline in its own right, scholars started to draw attention to a special status of the translated language. Toury (1979) referred to the product of translation as a *separate linguistic system* or *interlanguage*. According to the linguist, the language used in translation tends to exhibit manifestations of the target language (TL) that he characterized as *translationese*. For Toury, the translated language is a compromise between the adequacy and the acceptability, where adequacy aspires to the full representation of the source language (SL) message, and acceptability aims at the appropriate position of the translation within the target language and culture. Toury saw the *interlanguage* produced in the result of translation as an idiolect specific to the translator or a group of translators, which allowed room for systematic studies of the interlanguage forms occurring in translation.

Duff (1981) identified translations as *the third language*, highlighting that a translated text can be considered as a coherent unit unless it is not a combination of several languages and styles. Frawley (1984) coined the term *the third code* in the literature, describing translationese as a unique sub-language in its own right that incorporates influences from both the source and target languages. The linguist claimed that translated language has a dual lineage as a result of the influence from both the source and the target languages. Shuttleworth and Cowie (1997) explain translationese as the frequent appearance in the target text of items of the source culture realia. Such items, or features, as the authors point out, do not pose a threat to the target language norms, but add an indefinable translated feel. For Shuttleworth and Cowie the term *third code* is used to describe subtler linguistic deviations from the target language standard. Thus, the third code can extend and enrich the linguistic repertoire of the target culture.

The same phenomenon of translationese also received pejorative interpretations: translationese was attributed to the poor competence of translators. For Baker (2019), translation is the outcome of the conflict between the source and the target codes, which requires a certain level of mastery to resolve. Ilisei (2012), although holding a non-evaluative view, justified such a stand with the possible influence of the conventions within literary translation according to which a good translation should not let itself be easily identified as such. For Rayson (2008, p.1) translationese are manifestations of "the skewed nature of translated texts" that in the best scenario give the flavor of a "mechanical or monotonous translated text", while also may result in "inaccurate and erroneous translation".

Gellerstam (1996), who is considered to be one of the first scholars to coin the term *translationese*, characterised such linguistic variations of translations as the *unmistakable fingerprints* of the source language revealed in the target text. In his work, the linguist

sought statistically significant discrepancies between original Swedish texts and Swedish texts derived from English translations. Although the terms *translationese* and *the third code* are closely related, Granger (2018) prefers to adopt Flawley's *the third code* in her cross-linguistic study of meta-discursive markers to avoid the pejorative connotation that was sometimes attributed to the term *translationese*.

Koppel and Ordan (2011) introduced a different concept to refer to the peculiarities of the translated language. The scholars spoke about the *dialects* of translationese by which they referred to translations from different source languages into the same target language. Koppel and Ordan (2011) assumed that in the result of translation, each source language produces its specific traits in the target language that may be viewed as a new language variant or a dialect.

Much of the recent and current research in the field of translationese has shifted the focus away from the evaluative statements about the professionalism of translators. Ilisei (2012, p.50) outlined translationese as "all the potential features that may prove to be specific to the translational language in contrast to non-translational language". Taking into account previous advances in the field (Gellerstam, 1996; Baker, 2019; Baroni and Bernardini, 2006; Volansky, Ordan, and Wintner, 2015), Chowdhury, España-Bonet, and Genabith (2020, p.1) define translationese as "systematic differences between translations and texts originally authored in the target language, in the same genre and style".

When defining translationese, Kunilovskaya and Corpas Pastor (2021, p.5) emphasise the "differences in frequencies of language items between translations and non-translations in the TL regardless of their hypothesised cause". In their work on the register variations in English-to-Russian translations, translationese is investigated in relation to the source language texts and non-translated target language texts across different registers. The present research will adopt the definitions of translationese by Ilisei (2012) and Kunilovskaya and Corpas Pastor (2021). We will discard the interpretation of translationese as poor work of translators and rather view this phenomenon as an inevitable aspect of translation that may reveal objective properties of translated texts and yield insights into the translation process.

2.1.2 Translationese Hypotheses

The early works in the descriptive translation studies distinguish two possible sources of translationese. The first line of thought, expressed in House and Blum-Kulka (1986), Baker (1995), Shlesinger (1989), Zellermayer (1990), Weissbrod (1992), and Baker (2019), supported the conviction that linguistic variations arise from the translation process independently of the source or the target language. Baker (2019) stressed the universal nature of the properties of translations by introducing the notion of translation universals that she characterised as "features which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems". Primarily, the translation theorist named four tendencies in translation which

she thought to be *universals*: explicitation, simplification, conventionalisation, and normalisation. *Explicitation* refers to the tendency of explaining in more detail the content of the source text. *Simplification* is associated with facilitating the comprehension of the translation by opting for transparent vocabulary and/or changing the structure of the source text in translation. *Conventionalisation* (or levelling-out) hypothesis suggests that translations are more similar to each other than their sources. *Normalisation* stands for the choice of very frequent lexical, stylistic, and structural elements items instead of a variety of less frequent options.

The phenomenon of universals is perceived by Toury (1995) from a different perspective with a distinction being made between the law of growing standardisation and the law of interference from the source language. The law of standardisation involves neutralising the unusual linguistic entrances of the source language by selecting equivalents in translation more typical to the target culture. Toury gave particular importance to the law of interference, placing it at the centre of translation studies. The law of interference, that had migrated into translation studies from research on second language acquisition, was introduced to describe how the source language structures dominate over the target language structures in translations. Toury's assumption stems from the premise that translationese is bound to be influenced by the source language, while such revelations of the SL norms should not be viewed as universal. The translation scholar questioned the universality of any statements, thus favouring the term *laws of translation* to group the instances of translational behaviour. As the linguist later explained it Toury (2004), unlike *translation universals*, *laws* have the potential of exception built into them that helps to avoid misunderstanding and makes it easier to experiment with.

Following the paradigms within previous research in translation studies (Frawley, 1984; Toury, 1995; Baker, 1995; Baker, 2019), Chesterman (2004) prefers to avoid the rhetoric of what translations should not be and discusses the possible sources that shape translations. To explain his hypothesis, the author singles out three types of texts involved in the process of translation: translated texts that are guided by the source texts on the one hand, and reference texts on the other hand. Reference data or *non-translations* comprise texts that were written in the same communicative situation as translations but originally in the target language. The translated texts always aspire to the fluency of the reference texts, the connection that Chesterman denoted as *the quality of naturalness*, *the relation of acceptability*, or *the relation of target text fit family*. Similarly, translations strive to preserve *the relation of equivalence* with the source text.

Chesterman's hypothesis distinguishes between two groups of translation universals: S-universals and T-universals. S-universals are tendencies in translation that are revealed through a comparison of source texts and their translations, whilst T-universals are properties that are identified through looking for linguistic variations between translated and non-translated texts. Possible S-universals according to Chesterman include lengthening, i.e. the tendency to make longer translations than their source texts; the law of inter-

ference (Toury, 1995); the law of standardisation (Toury, 1995); normalisation (Baker, 1995); reduction of complex narrative voices; explicitation (House and Blum-Kulka, 1986; Klaudy, 1996); sanitation, i.e. the usage of more traditional collocations (Kenny, 1998); retranslation, i.e. is a new translation of a previous text that has already been translated that results in a possible loss on the equivalence; reduction of repetition (Baker, 1993). The second group outlined by the linguist, that of T-universals, embraces simplification i.e. (Baker, 1996); (Laviosa-Braithwaite, 1996); conventionalisation (Baker, 2019); untypical and lexical patterning (Mauranen and Olohan, 2000); under-representation of specific items in the target language (Mauranen and Olohan, 2000).

Several works on the special character of translations revealed confusion as to which translation indicators are associated with which hypotheses. For example, Puurtinen (2003) looked into the frequency of complex non-finite constructions as found in translated and non-translated Finish texts that were sourced from childrens literature. Initially, the scholar hypothesized that translations would show simplification, explicitation, normalisation and concretisation. The experiment revealed that the frequency of non-finite constructions increased in translated texts, which does not support simplification and explicitation hypotheses because non-finite constructions tend to complicate the comprehension of the text. The normalisation hypothesis was refuted as well: Puurtinen observed that a high number of non-finite contractions is not typical of the syntactic norms manifested in the originally authored Finish children literature. The scholar concluded that non-finiteness should be considered as a feature of translationese in Finish childrens literature.

Because of the controversial nature of the *universals*, in the subsequent studies, the focus was shifted to the experimental evidence of certain trends in language use in translated texts. Bernardini and Zanettin (2004) saw *the quest for universals* as pointless. In their work that aimed at designing a bidirectional translation corpus, the scholars emphasise that measuring the frequency of certain linguistic features in translation – such as type-to-token ratio, lexical density, and other structural and lexical features – contributes to a more fruitful discussion than searching for universals. Volansky, Ordan, and Wintner (2015) refrain from referring to translation universals in favour of adopting the term *features* – a choice which the scholars support with several observations. The authors put the emphasis primarily on data and empirical findings, rather than translation theory itself. The features are seen as pertaining to more than one translationese hypothesis. In their experiment, Volansky, Ordan and Wintner (2015, p.2) reveal that some features – such as mean sentence length – depend on the source language, which, according to the authors, undermines the "universality of universals". Finally, the choice of the term *features* was motivated by a connection to the machine learning methodology used in the experiment. Gradually, the concept of features has gained prominence in research on translationese. Current studies either explore the instances of translationese in general or seek to attribute the characteristics of translated language to one or more translationese

trends.

2.1.3 Translationese Indicators

A feature can be defined as a unique property or characteristic of a phenomenon that can be measured (Bishop, 2006). In our study, under features, we understand linguistic patterns that serve for automatic analysis of translated language. Translationese was first investigated in terms of grammatical patterning. Gellestam (1986) – one of the first scholars who brought into academic discussion the phenomenon of translationese – contrasted novels translated from English into Swedish with novels originally authored in Swedish. The experiment revealed that translated texts contained a higher number of *adjective-noun* patterns and a different proportion of relative clauses found in translations and in non-translations. Laviosa (1998) used *lexical density*, i.e. the proportion of content words to non-content words, and *mean sentence length* to see if linguistic properties of translated newspaper articles were also seen in translated narrative prose in English.

Santos (1995) approached translationese from the grammatical point of view. For her study, the linguist compiled a bidirectional translation corpus of English and Portuguese and selected *the frequency of the progressive constructions* and *the occurrences of the present perfect and the complex aspectual classes that may require partitioning in the process of translation.* It was found that progressive constructions are more frequent in English than in Portuguese; this was proved by a higher frequency of progressives in Portuguese texts translated from English.Although the study had some limitations – the author pointed out that the bidirectional translational corpus of English and Portuguese texts was small and unbalanced –, it contributed to broadening the panorama for features design in research on translationese.

Olohan (2001) experimented with syntactic and lexical features on significantly larger corpora: The Translational English Corpus (TEC), containing translations into English from a wide range of languages, and a subcorpus of the British National Corpus (BNC) that served as comparable data. Based on the premise that translators strive to make the final product more explicit by favouring the use of optional syntactic features even when it is not required by the language norm, the author elaborated a set of features that could be omitted in English translation. The syntactic indicators of explicitation for the experiment included the omissions of *subject NP*; *complementiser that*; *relative pronoun wh-/that*; *to be from complement clause; predicate*; *modal should from a THAT complement*; *preposition before complementisers that, for and to*; *complementiser to*; *after/while in (after) having and (while) *ing*; *in order.* It is a priori impossible to trace omission by its occurrence, that is why the established methodology was reversed to compare the instances of the given features in both corpora.

The concept of features was fully adapted to research on translationese using the machine learning methodology. Inspired by studies of supervised machine learning for text classification, Baroni and Bernardini (2006) opened up a new line of research on trans-

lationese. On the corpus of translated and non-translated journal articles in Italian, the scholars demonstrated how machine learning methodology, in particular Support Vector Machines (SVM), could be applied to distinguish translations from non-translations. Baroni and Bernardini used a combination of surface lexical and syntactic features, experimenting with the size (unigrams, bigrams, trigrams) and type (wordform, lemma, pos tag, and mixed) of language units, to classify the texts as translations and originals. The linguists created both weighted and unweighted feature sets to represent each text from the corpus as a feature vector. The SVM algorithms showed better classification results using either of the following representations 1,2,3 than using 4, based on unigram wordforms, unigram mixed representation, and bigram mixed representation, while the combinations with trigrams archieved lower accuracy.

To search for specific characteristics of translated texts, Baroni and Bernardini (2006) singled out four classes of features: non-clitic personal pronouns, adverbs, punctuation marks and non-finite forms of verbs. Each feature class under investigation was excluded from the training model one by one. Non-clitic personal pronouns were more frequent in translations, while adverbial constructions decreased in frequency in translated journal articles. It was difficult for the authors to interpret the effect of removing punctuation marks and non-finites from the model because the differences between the models with them and without them were not statistically significant. The results from training the model without adverbs and non-clinic personal pronouns suggested that these feature classes could be valid indicators of translationese. The authors concluded that the designed set of features was rather exploratory and required to be further investigated and tested on more language pairs.

Similarly, Ilisei (2010) investigated manifestations of translationese in Romanian newspaper texts using machine learning approach. Ilisei (2010) proposed thirty-eight language-independent features that were used to train the model to distinguish translations from non-translations. The linguistic indicators were grouped into two sets. The first set included the features that reflect translationese in general, such as *grammatical words and categories*, *lexical words and categories*, *proper nouns* and other. The second class accommodated the features that were attributed to the simplification hypothesis, such as *lexical richness*, *mean sentence length*, *word length*, *the number of simple sentences as normalised by the total number of sentences*, *information load*.

Volansky, Ordan and Wintner (2015) used text classification methods to compute the differences between original texts in English and translations from 10 European languages into English. Taking into account previous discoveries in the field, the scholars elaborated a set of 32 features which they group around four translationese hypothesis, namely explicitation, simplification, normalization, interference, and a miscellaneous class. While determining the feature set, the authors stressed that the features must:

1. represent common language properties that we would expect to find in both types of texts

2. be content-independent, i.e. the differences between the texts must not be influenced by the domain or genre

3. be easy to analyse in terms of the differences between translations and non-translations that they indicate

Volansky, Ordan and Wintner (2015) were more concerned with the methodological contribution of their work rather than the accuracy of the outcomes; therefore, each feature is described in detail to allow for the replication of their findings and further investigation by other researchers. The authors assigned each feature to a particular translationese trend. The simplification hypothesis was explained through *lexical variety*, *mean word length* (in characters), *syllable ratio*, *lexical density*, *mean sentence length*, *mean word rank*, *most frequent words*. Explicitation was associated in terms of *explicit naming*, *single naming*, *mean multiple naming*, *cohesive markers*. Normalisation hypothesis was modeled by *repetitions*, *contractions*, *average PMI* (Pointwise mutual information), *threshold PMI*. Interference was analysed through *POS n-grams*, *character n-grams*, *prefixes and suffixes*, *contextual function words*, *positional token frequency*. Importantly, they ran separate classifiers for each of their features to determine their effectiveness, which makes their analysis univariate. The most effective features that yielded the classification accuracy in the area of 100-66% being used as a single input to a classifier include (in the order of importance): Contextual function words, POS trigrams, Function words, Positional token frequency, Cohesive Markers, Mean word rank, TTR, Threshold PMI.

In addition to computing the characteristic features of translated and non-translated texts through putting them in a shared vector space, Evert and Neumann (2017) suggested visual representation of the results via scatter plots. The authors used machine learning techniques such as Linear discriminant analysis (LDA) and Principal component analysis (PCA) to reduce the dimensionality of the feature space to a 2D image (the study also provided 3D visualisation of the computed data) that could be visually analysed. To them, new methodological solutions made it possible to depart from the univariate analysis, i.e. picking a single feature to analyse the translated language, and adopt multivariate analysis, i.e. including as many linguistic indicators of translationese as possible, because they used features shared by the source and target languages in their experiment. Another important innovation that this work introduces into the field, is including automatic analysis of source texts into the research design. To trace the difference between translated and original language on a bidirectional English and German corpus, the scholars designed 27 lexico-grammatical features.

For their study on the peculiarities of translated Russian across registers, Kunilovskaya and Corpas (2021) came up with an extensive set of linguistic indicators of translationese that ware extracted from the Universal Dependencies (UD) annotation (Straka and Straková, 2017) and a predefined list of features. Excluding the features extracted from UD annotations, Kunilovskaya and Corpas (2021) came up with several novel mea-

sures for English and Russian language pair comprising features based on pre-defined lists of connectives and discourse markers for English and Russian. The authors distinguished between additive, adversative, causative, temporal sequential discourse markers.

Their study presents a comparison between source, target and reference texts across four registers: general media, popular science, fiction, news commentary. The scholars ran a SVMs classifier on the full feature set and then excluding morphosyntactic or lexical indicators in turns. Translationese classifications (translations vs non-translations) showed the accuracy of over 95% for all registers except mass-media texts, with still good performance of 87%. It was found that morphosyntactic features were better predictors of translationese than abstract lexical features (such as ratios of lemmatised n-grams in the top and bottom frequency quatiles or language model perplexities). When lexical features were removed, the performance of the classifier was one to two percent less accurate; yet, with lexical features alone the classification accuracy deteriorated from 7% for news commentary to 17% for popular science. The authors attributed the higher prominence of morphosyntactic translationese indicators in popular science to less variability of non-translations, and probably to greater divergence of cross-linguistic conventions in this register.

2.2 Methodological Approaches to Translationese

2.2.1 Manual, Semi-Automatic and Automatic Analysis

Manual analysis, sometimes enhanced by simple corpus-based techniques, was the mainstay of early comparative studies between originally-authored texts in the TL and translations. For example, Santos (1995) relies on identifying repeated instances of progressive constructions, present perfect tense and complex aspectual classes in a relatively small corpus of English and Portuguese texts. While this research paved the way for the incorporation of grammatical features in experiments on translationese, most of the analysis was done manually – based on the observations of the parallel concordances extracted from the corpus – which allowed to trace only a few grammatical phenomena of translated language. In a series of investigations into complex non-finite constructions in translated and non-translated Finish childrens literature, Puurtinen (2003) starts from the manual syntactic examination of the texts and further sharpens her research with improved corpus methods that allowed to operate on a larger amount of data Puurtinen (1998). Due to the constraints of manual analysis, each study mentioned above only investigated a small number of features.

Advancements in technology – in particular, the invention of more sophisticated computer software and the ability to access large amounts of electronically stored data – were gradually refining the methodology for identification of translationese. Laviosa-Braithwaite (1996) used statistical measures as lexical density, sentence length, typeto-token ratio to find distinctive features of translations in newspaper subcorpora from the

Translational English Corpus (TEC) and the British National Corpus. In favour of the simplification hypothesis, the experiment showed that all the indicators were lower in translated texts, although the methodology did not give a solid foundation for proving this claim. Baker (2004), a translationese theoretic and one of the first scholars who advocated corpus-based approach to the study of language variations in translation, experimented on the same corpora to trace lexical patterns and their manifestations in translated language. The scholar realistically reported on the existing methodological issues within translationese research. As an illustration, the software available at that time did not allow to search for all the n-grams with the nucleus "event", but just to specify a number of words in the search: 3-word pattern would not capture "in the event that", and 4-word pattern would skip "in any event". The list of the candidates generated by the software was not reliable, so further manual corrections were made.

Machine learning brought significant changes in methodological approaches to the study of linguistic variation in translation. Following the paper by Baroni and Bernardini (2006), SVM was repeatedly shown to be one of the most effective ML algorithms for the translationese detection task. Under this method, each text is represented as vectors of features. Based on the features, the machine learning algorithms find the border between the vectors that belong to one class of texts or another, i.e. classify the texts into different groups, in this case translations and non-translations. As believed by the authors, this methodology allows to make full use of the information stored in the texts which is impossible with single feature extraction. Both weighted and unweighted frequency vectors were calculated for each feature set. For weighting the features, *tf-idf* (term frequency-inverse document frequency) algorithms was used, i.e. the frequency of each feature in an document divided by the number of documents in which the feature appears.

There are several decision rules for the machine learning classifier to determine if the output belongs to a one or the other class. Majority voting method, and this is clear from the actual name of the term, categorises a text as a translation only if the majority of the classifiers indicates it as so. Recall maximization method labels a text as a translation when at least one classifier decides on it. Baroni and Bernardini (2006), as claimed by the authors, were the first who apart from majority voting experimented with recall maximization, with the later giving more fruitful results in terms of all the four indicators set by the authors: 85% accuracy, 80.9% precision, 91.7% recall and 85.9% F-score, i.e. precision and recall combined in one number. The experimental data revealed that high quality translations share enough linguistic characteristics to be recognised by the model with precision of 90% and recall above 80%. Concluding on the performance of the best combinations, the scholars report that the SVMs algorithm relies more heavily on grammatical/syntactic features to classify texts as translations, than on lexical indicators.

Following a similar methodological approach, Ilisei et al. (2010) trained a machine learning model to classify translations and non-translations from monolingual comparable corpora of medical and technical texts in Spanish. Ilisei tried a range of algorithms,

including Jrip, Decision Tree, Naive Bayes, BayesNet, SVM, Simple Logistic and Meta-classifier. Their results confirm that SVM is the best for the translationese classification task. The bag-of-words features, under which grammatical, structural and contextual information is ignored and each word is assigned a number according to the number of times it appears in the text, was not applied to prevent the model from classifying the texts according to their topics. The linguists obtained the accuracy of 87.16% on the categorisation task and 97.62% for separate samples from the technical domain.

In the subsequent research, I. Ilisei and Inkpen (2011) dealt with the classifying task in a slightly refined manner and tested it on a parallel corpus of Romanian newspaper texts. Importantly, the corpus comprised articles translated from different languages, which excluded the possibility of the classifier tracing a particular source-language-specific translationese. For their experiment, the linguists added more features to their set and reduced the number of classifiers as compared to I. Ilisei and Inkpen (2011). All the classifiers showed remarkable accuracy between 91.71% for Naive Bayes and 98.90% for SVMs by which the researchers proved that SVM works best for the detection of translationese even for translations for a number of source languages.

Chowdhury, España-Bonet, and Genabith (2020) report results of a study which is representative of a relatively new direction in translationese studies: investigating the impact of the source language on the properties of translations. They assessed translationese effects in English translations from multiple SLs from EuroParl collection by (Koehn et al., 2005). Translations and comparable texts in the TL were compared to trace the distances between texts translated from languages belonging to different families and non-translations. For this aim, Chowdhury, España-Bonet, and Genabith (2020) measured the degree of isomorphism between embedding-based semantic spaces, i.e. checking to what extent translations and originally-authored texts in the target language generate similar spaces and whether the divergences between those spaces were more pronounced depending on the language family of the source language.

The initial hypothesis was that the embedding spaces would be more isomorphic for translations from the languages that belong to the same or closely related language family to English. The vector spaces for translations and non-translations were created with fastText Bojanowski et al. (2017) from texts represented as sequences of part-of-speech tags, semantic tags, and synsets from WordNet Miller (1998). Eigenvector Similiarity (EV) metrics was used to compute the isomorphism between texts translated into English and English originally-authored texts. Based on the EV distance score, it was confirmed that translations from Germanic family were closer to English original data, followed by Romanic group, while the Balto-Slavic group was situated the furthest in the semantic space. The experiment confirmed the well-established belief that a lesser degree of interference should be expected in translations from languages belonging to the same or similar language family which is explained by higher semantic isomorphism.

2.2.2 Machine Learning Methods

Text categorisation methods lie at the boundary of machine learning and information retrieval. Accordingly, automatic classification can be based on a predefined classification scheme and an already existing set of classified (annotated) documents. The classification algorithm works on a set of documents divided into a training and a test collection; they are also commonly referred to as the training set and the test set. Obviously, the training set has to be larger than the test set in order to get a higher machine learning results. The classification lies in predicting the labeled class to which a given document belongs. Each document receives a set of attributes or features. A classification algorithm is then applied to select the documents that best fit the given class, given the features, or, in other words, contain the feature patterns that are associated with a particular class.

A vast majority of methods for text categorisation are based on the assumption that documents belonging to the same category (class) have the same statistical properties. In Natural Language Processing (NLP) tasks these properties are similar frequencies of specific linguistic items. As an outcome of a classification task, each document is assigned to a particular category with a certain degree of probability. A number of powerful approaches, methods and algorithms were developed within the ML paradigm: *Artificial Neural Network (ANN)*, *Support Vector Machine (SVM)*, *Classification and Regression Tree (CRT)*, *Regression and Classification Tree Ensembles (Random Forest, RF; Gradient Boosting, XGBoost)*, *Deep Neural Networks and Deep Learning (DL)*, *Kernel Methods*, and many other. Some of these methods were applied in previous studies on translationese (Baroni and Bernardini, 2006; Ilisei et al., 2010; Ilisei, 2012).

Decision trees, for instance, partition data into groups based on the values of the feature space variables, generate a hierarchy of "if-then" operators that classify the data. In order to decide which category a given document belongs to, it is necessary to go through the questions at the nodes of this tree, starting from its root. If the answer is positive, the algorithms proceeds to the right-hand node of the tree, if negative, it goes to the left-hand node. The next question is related to the corresponding node. To determine the variable with maximum classifying power, the criterion of the information weight of the word in the heading is used. Usually after constructing an exact decision tree, various tree truncation and transformation procedures are applied to the resulting tree in order to balance tree complexity (number of nodes) and learning quality. A classic approach to transforming decision trees is the C4.5 algorithm.

One of the main drawbacks of the decision tree method for text classification tasks is that the decision tree algorithm gives equal weight to positive and negative branches in the nodes. A large number of negative branches in a label description can lead to difficult to interpret rules and over-fitting of the classification algorithm. The advantages of this method are that the tree can be easily analysed and the result of the algorithm can be interpreted in clear terms. Moreover, there are programmes for graphical representation

of decision trees.

Naive Bayes classifier is usually used in text classification tasks such as spam filtering, automatic categorisation or determining the tonalities of a document. This method applies a probabilistic model in which classification and inclusion in the appropriate document category is done by estimating the probability of occurrence of words in the document. The probabilities can be used to estimate the closest categories of the test document.

The main advantages of a Naive Bayes classifier are its simplicity of implementation and the low computational cost of training and classification. In those rare cases where the features are truly independent, the naive Bayesian classifier is optimal. The main disadvantage of the method is its relatively poor classification quality in most real-world problems. This method is often used as a baseline method when comparing different machine learning methods.

The k-nearest neighbors (KNN) algorithm is one of the most studied and highly accurate algorithms used in creating automatic classifiers. It was first proposed back in 1952 for solving problems of discriminant analysis. In studies analyzing the performance of various machine learning algorithms for the task of text classification, this method performs with some of the best results (Yang and X. Liu, 1999). The method is based on the idea of finding and categorizing the most similar documents to the text under analysis and, based on the knowledge of their categorical affiliation, classifying the unknown document. In order to determine the label relevant to a particular document, this document is compared with all the documents in the training sample.

For each document in the training sample, the cosine of the angle between the feature vectors is calculated and assigned to it, taking into account its relevance to the closest documents. Categories with relevance above some given threshold are considered relevant to the document. The parameter k is usually chosen between 0 and 100. If the category is monothematic, the class with the maximum value is selected. If a document can be assigned to more than one category, which is the case of multiclass categorisation, the classes are considered relevant if the value exceeds some predefined threshold.

The main feature that distinguishes this method from others is that it does not have a learning phase. In other words, the document belonging to a category is determined without constructing a classification function. The main advantage of this approach is the possibility to update the training sample without retraining the classifier. This property can be useful, for example, when a training collection is frequently replenished with new documents and retraining takes too much time. The classical algorithm proposes to compare the analysed document with all the documents in the training sample and therefore the main drawback of the described method is the long running time of the classifier during the classification phase.

Artificial neural networks have been widely studied in the field of artificial intelligence for data analysis since 1986. They represent a mathematical model with its software or hardware implementations based on the similarity of nerve cell networks of a living

organism. Neural networks are one of the best known and oldest methods of machine learning.

Artificial neural networks are an adaptive system consisting of a group of connected artificial neurons. The system can be trained to change internal states, mapping document relationships and document categories. In order to classify texts efficiently, it is necessary to find a rational structure and topology of the neural network. Known architectures of neural networks include single layer (or multilayer) perceptron, neural network Gaussian classifier, Kohonen network, embedded propagation network, cascade network (Yang and X. Liu, 1999). All of the above topologies are highly accurate in processing both linear and non-linear data.

The Support Vector Machines (SVM) method is a solution for general purpose pattern identification. It represents a process of creating a hyperplane that can separate positive and negative examples in a multidimensional feature space, where training documents are represented as vectors. The SVM training algorithm then builds a model that assigns new samples to one or the other category. The SVM method was developed by Vladimir Vapnik in 1995 (Cortes and Vapnik, 1995). It was first applied to the text classification problem by Thorsten Jochims (Joachims et al., 1999). In its original form, the algorithm solved the problem of distinguishing between objects of two classes. The method has become very popular due to its high efficiency. The approach proposed by V. Vapnik for determining which of the two predetermined classes the analyzed pattern should belong to is based on the principle of structural risk minimization. It is usually referred to as a binary linear classifier, it could also be adapted for classification of multiple items.

The SVM model is designed in such a way that samples from individual categories are separated by a wide blank space. It is clear that a separating surface (called hyperplane) that runs through the middle of a strip separating two classes works better than a separating surface that lies very close to instances of one or both classes. Some classifiers find at least one linear separator, others, such as the Naive Bayes method, find the best linear separator using a particular criterion. In contrast, the SVM classifier looks for a separating surface that is situated as far away as possible from any data points. The distance between this surface and the nearest data point is called the classifier gap. The support vector machines method necessarily implies that the solver function is well defined by a small subset of the data that affects the position of the hyperplane. These points are called reference vectors because, in vector space, a point can be viewed as a vector between the origin and this point. Points that are located closer to the separating surface might generate vague results. That is why a wider gap between the two classes of the SVM classifier works better for maximising the accuracy of the classification.

2.2.3 Language Type Involved

Since the language of translation was recognised as a separate language type or a third code Frawley (1984) scholars started to search for the possible sources that influenced such non-

standard language use. Different hypotheses about the universal characteristics Klaudy (1996), Baker (1995), Mauranen and Olohan (2000), and Olohan (2001) or the laws of translation Toury (1995) began to emerge, as discussed in previous sections. The proposed claims of early translationese researchers were later roughly divided into S-universals, or those related to the source language (SL), and T-universals that were associated with the influence of the target language (TL), as put together by Chesterman (2004). It follows from these observations that changes in the linguistic behaviour of a translator can be effected by the norms of the source or the target language. Thus, we can differentiate between two types of corpora that are used for research on translationese: parallel and comparable. A parallel corpus is composed of texts in one language and their translations in another language, while a comparable corpus is a set of texts representing one language. The comparable corpora used for translationese detection usually contain texts produced in the same communicative situation as the translations but are originally authored in the target language.

Early descriptive studies mainly approached translationese through the comparison between the translations (T) and originally-authored texts in the target language (O). Yet, the pioneering works in the field do not show a clear position on whether the source texts (S) should be taken into account when analysing translationese. A study conducted by Gellerstam (1996) examined language deviations between Swedish literature (translated from English) and original Swedish literature. Although the author was comparing translations to the Swedish language norms, he did not exclude the possibility that some of the translationese pertained to the source language. Bakers (1995, 2019) frame of reference for her theory of 'universals' was the language usage in translations as compared to those in the target language. Laviosa (1997) and Laviosa (1998) exploited translations and non-translations comprising the English Comparable Corpus (ECC) to study the distinguishing patterns of translational English prose. Toury (1995) translations between the two types of texts to formulate the law of interference as the dominance of the source language and the law of standardisation as the adjustment to the target language rules, although his mainstay was the comparison between the translated texts and originally written texts in the target language. Mauranen and Olohan (2000) stressed the importance of a parallel corpus to determine whether a target text feature follows a given source language feature leading to interference.

The emblematic work of Baroni and Bernardini (2006) significantly extended the scale of research into translationese by introducing new methodological perspectives that allowed to place different types of texts in a shared vector space through a set of features. For their experiment, the scholars compiled a corpus of translated and non-translated Italian articles from geopolitical domain. Different combinations of features with SVMs algorithm were tested to place translated texts against comparable texts in the target language. Some researchers followed this model and built their experiments around translations and original texts in the target language. Ilisei et al. (2010) trained several machine

learning classifiers to see if they can distinguish translated and non-translated texts from medical and technical domain while testing the simplification hypothesis on the same corpora. The classifiers obtained high accuracy for the classification task when tested on simplification features and lowered in performance when the features were excluded.

Popescu (2011) demonstrated that translationese could be traced using machine learning algorithms that function at the character level. The SVMs algorithm was posed a binary classification task of distinguishing between novels written by English and American authors and literary translations from French and German into English. The first experiment was conducted on the entire corpus and yielded almost 100% accuracy of the learning method, which resulted suspicious to the authors. Contrary to the core principles of machine learning, in the second setting the authors experimented with training the classifier on the translations from French and target language texts written by British authors and then testing the model on the translations from German and the originals by American authors. The second attempt rendered 46% of accuracy which gave the impression of randomness. For the third experiment, originally authored novels in French were added to the corpus to detect the factors that might cause confusion of the classifier. Thus, the sub-strings referring to French proper names were eliminated and this substantially improved the performance of the classifier even on the training and the test data from different source languages. As a side note, the authors considered whether it was appropriate to use a reference corpus of the source language in translationese related experiments.

Evert and Neumann (2017) argue that it is not feasible to study the divergences between translations and non-translations without involving the source text. The scholars highlight that the source language and register diversity are among other key factors that come into play when studying translationese. In the reasoning, the linguists refer to notion of the shining-through effect introduced by Teich (2003), i.e. the prevalence in translated text of features that are more proper to the source text, given that such features exist in both language systems, and build their study around the source language-induced variations of translations. For the experiment Evert and Neumann (2017) collected a bidirectional translation corpus of English and German texts to explore how translations differ from comparable target texts in terms of translation direction and other variables such as the register. Importantly, they threw their findings on translationese into the perspective of the source texts. As a result, a distinction was made between the authentic shining-through characteristics derived from the source language and text-specific shining-through that can be traced back to the style of the author or the domain. A set of multiple linguistic features was used to first compute the differences between the text types and then to visualize the results using machine learning algorithms. The scatter plots revealed that English originals and translations into English had almost equal distributions, but translations into German were moved significantly away from German originals. Those experimental results proved that there was more source language shining-through effect

in English-to-German than in German-to-English translation. The authors attributed the difference to the language prestige effect.

2.2.4 Bottom-up and Top-down Approaches

At different stages of development in the field of translation studies, scholars adopted top-down and bottom-up reasoning to explore the phenomenon of translationese. These well-established concepts refer to two fundamentally different approaches to the study of translated language. Bottom-up method starts with identifying smaller units, i.e. grammatical, syntactic, lexical or stylistic manifestations of translated texts, and moves upwards to group some of these traits under one hypothesis or another. Top-down analysis, on the contrary, starts with a hypothesis (for example, if translations have lower mean sentence length than non-translations, it will be interpreted as a sign of simplification) and experimentally confirms or disproves it. Thus, Gellerstam (1996), the author of the extensively cited early work on translationese, started from the observations of the data to arrive at some generalizations about Swedish texts translated from English. The linguist searched for statistically significant divergences – with the frequency over 100 – between original Swedish texts and translations from English and arrived at the conclusion that translations contained less colloquialisms, more standard vocabulary and loan words, among other features.

In the bottom-up line of research, the empirical results of some early studies on translationese led the scholars to propose different hypothesis on the nature of translated language. Klaudy (1996) and House and Blum-Kulka (1986) found that translated language was more cohesive and therefore translations may all share the trait of explicitness. Laviosa-Braithwaite (1996) proposed simplification hypothesis after revealing that translated texts contained less varied vocabulary, lower lexical density, and larger proportion of high-frequency items. Mauranen and Olohan (2000) observed "untypical lexical patterning" in translation which lead to the creation of the like-named hypothesis.

In contrast to the bottom-up methods of some pioneers of translation studies, Santos (1995) created several candidates for indicators of grammatical translationese to be found in a parallel corpus of English and Portuguese texts. The linguist reported that although the data collected for the experiment did not allow for a large-scale investigation, a few translationese effects associated with the influence from the source language were found. Olohan (2001) conducted a broader study in terms of the corpora and tested the explicitation hypothesis on translated texts into English and originally authored English texts. The data was analysed from the syntactic standpoint to find evidence supporting the claim that optional syntactic elements may be used more frequently in translated texts than in originally written texts in the same language. While some of the indications proved the initial explicitation claim concerning the language of translations, any generalisation could not be drawn based on the methodology employed because of the statistically insignificant results for certain syntactic indicators.

Corpas et al. (2008) designed a set of lexical and stylistic features to validate the simplification and convergence, i.e. the assumption that translated texts share similarity traits, hypotheses on a large corpus of medical and technical texts. The simplification hypothesis was tested against the source texts and involved lexical features, such as lexical density and lexical richness, and stylistic features such as sentence length, the proportion between simple and complex sentences, the frequency of discourse markers and readability, which was computed on the basis of Automated Readability Index, Coleman-Liau Index, and Flesch-Kincaid Grade Level Readibility Test. To address the convergence hypothesis, comparable non-translations from the same genres were used and apart from computing the possible lexical and stylistic features of translations mentioned above, the experiment embraced syntactic features. Specifically, the two corpora were part-of-speech tagged and represented as frequency vectors of 3-grams using cosine similarity and recurrence metrics. The preliminary hypothesis of the authors that simplification represented through the designed lexical and stylistic features can be found in translations was confirmed for the technical corpus, but not fully validated for the medical texts. Conversely, no evidence for the convergence hypothesis based on the selected feature set was found. The scholars conclude on the need to explore more features; in particular, the indicators of idioms and multi-word expressions.

Lapshinova-Koltunski (2015) conducted a study on linguistic characteristics of several translation variations in which she aimed at validating simplification, explicitation, normalisation as opposed to shining-through effect, and convergence on English and German translations. The simplification hypothesis was tested through lexical density and type-to-token ratio. To the explicitation hypothesis the scholar attributed higher ratio of function words, specific terms used instead of more general vocabulary, disambiguation of pronouns, a higher number of cohesive elements, and an increase in nominal constructions as found in translations. The author pointed out that the features of simplification and explicitation are interrelated and may be diametrically opposed. Normalisation, i.e. the overuse of patterns that are prevalent in the target language, is opposed to shining-through, i.e., carrying over the pattern specific to the source language. The linguist expected a larger number of conventional collocations and neutralised metaphors in translations to verify normalisation, or prove the shining-through effect. According to the convergence phenomenon, the lexical, grammatical, and syntactic features should exhibit less variations in translations than in originals. The experiment revealed that lexical density and type-to-token ratio alone are not reliable indicators of simplification as well as the features selected to test explicitation, normalisation and shining-through that showed partly good results; while convergence hypothesis was proved on the selected feature set.

Recent research in the field of translationese is gearing again towards the bottom-up approach, enhanced with new solutions for methodology and feature set designs. The reason for this is the confusion as to which features are grouped under which hypothesis or whether a definite feature set is representative of a particular translationese claim

or a translated register. Kunilovskaya and Corpas Pastor (2021) drew attention to the mixed and contradictory results of some studies that sought to prove the universality of translationese hypotheses. The universality might be blurred by the domain, register, language pair and the feature set involved in an experiment, as resumed by the linguists. Kunilovskaya and Corpas Pastor (2021) agree with Zanettin (2017), who stated that translators style and extra-linguistic knowledge should be taken into account when generalising on the nature of translated texts, on the fact that enclosing certain linguistic indicators of translations within general hypothesis leads to controversial outcomes. Besides, some studies did not make explicit the methodology for extracting the features or the attribution of the results to such hypotheses as explicitation, normalisation or source language interference (Castagnoli, 2009; Olohan, 2001; Kunilovskaya and Kutuzov, 2017) cited in Kunilovskaya and Corpas Pastor (2021).

Aware of the controversies behind translationese hypotheses, Kunilovskaya and Corpas Pastor (2021) preferred to use the terms translationese effects by which they group similarities of the frequency patterns of linguistic features expressed in translations as compared to the source texts and comparable originally-written texts in the target language. For their multifaceted study of translation variations on English-to-Russian language pair, the scholars design a set of morphological, syntactic, and text-level features to explore the translated language in relation to the source and the target languages across four registers. The data suggested that fiction texts contained the smallest number of shining-through indicators, and the biggest amount of over-normalised and adapted elements. News commentary texts presented the highest number of anglicisation and over-normalisation in terms of features, which places the translations in this register between the source and the target languages. Shining-through was the prevailing translationese effect among all the registers with the highest proportion of features found in popular science texts. Apart from the prominent shining-through effect, mass-media texts showed a significant proportion of adaptation and anglicisation, which could be attributed to the tendency to copy the source language structures in translations.

2.3 Register in Translation-oriented Studies

2.3.1 Definitions of a Register

The literature on stylistics and sociolinguistics accommodates several concepts associated with the way the same language is used, depending on the social occasion, purpose, or interlocutor. Multiple perspectives of text analysis brought into the academic discussion such concepts as style, genre, or register that have undergone terminological fuzziness. One attempt to disentangle their uses is offered in Lee (2001). The genre is more closely connected with culture and social purposes behind the language use (Lee, 2001). The style and the register share more similarities as to the linguistic components investigated. But if the style focuses on the aesthetic characteristics of a text, such as an individual writing

style, the register integrates the analysis of linguistic peculiarities that are common for a certain text variety with the analysis of the context in which this text type is used. Our study will stick to Biber and Conrad's (2019, p.2) definition of a register as a language "variety associated with a particular situation of use". The linguists highlight two main components that frame a register which is a situational context and linguistic indicators of a particular text type together with a description of how those indicators operate within that particular context of language use.

Almost any type of text possesses its characteristics or, as language scholars refer to it, linguistic features that reflect the circumstances in which these language indicators are manifested and/or the purposes for which they are used. In his definition, Ferguson (1977) emphasizes the role of features as a distinguishing trait of a register: "A register in a given language and given speech community is defined by the uses for which it is appropriate and by a set of structural features which differentiate it from the other registers in the total repertory of the community". A. M. Zwicky and A. D. Zwicky (2015) characterised a register by 1) exclusion of certain linguistic features – for example, contracted forms are unacceptable in English formal writing; 2) special freedom concerning other features – for instance, copula verbs are often omitted in newspaper headlines; or 3) preference towards a certain class of features, as for adjectives in creative writing, as an example. A classic illustration of a register label would be *baby-talk* as first mentioned in Ferguson (1977). Zwicky and Zwicky point that there are also features common for many registers, as subject-verb agreement and modifier-noun order, although these features are language-dependent.

In computational linguistics and natural language processing, the theoretical notion of a register is often used interchangeably with a broader concept of a text type by which researchers understand a group of texts that share a set of computationally quantifiable linguistic characteristics. Machine analysis of textual corpora derives from the consideration that different communicative contexts require different linguistics means. Texts of the same register may share syntactic, lexical or grammatical features that can be computed and represented in a multidimensional space. One of the main purposes of computational register studies is to find characteristic traits of a text belonging to a particular register or to establish the relation between various registers. In research on translationese, register might be yet another important dimension of text analysis that could bring insights about the translation process. The task of a translationese researcher who undertakes a study on register variation would be to find how and to what extent different registers affect the product of translation.

2.3.2 Register-oriented Translationese Studies

There is a belief that the source and target languages are only two of many factors that influence the process of translation. Some of the earlier approaches to translationese mostly sidestepped the impact of functional language varieties on translation and focused

on the characteristic traits of translated texts regardless of register (House and Blum-Kulka, 1986; Klaudy, 1996; Baker, 1995; Baker, 1996; Baker, 2019; Laviosa-Braithwaite, 1996; Toury, 1995; Mauranen and Olohan, 2000; Baroni and Bernardini, 2006) and many others. Several studies of the last decade incorporated register diversification in the analysis of translated language.

Delaere, De Sutter, and Plevoets (2012) examined whether translated and non-translated Belgian Dutch employed standard language differently and whether the variations between translated and non-translated language are based on the functional text type or are derived from the source language. The profile-based chi-square distance was used to compute the distances between nine language varieties altogether: fiction, non-fiction, journalistic texts, instructive texts, administrative texts, external communication, Belgian Dutch translated from English, Belgian Dutch translated from French, Non-translated Belgian Dutch. It was demonstrated that both source language and text type influence the usage of the standard lexicon. The scholars found that fiction, non-fiction, and journalistic texts were more conventional in terms of lexical choice than other genres. The experiment revealed that Belgian Dutch contained more instances of standard language than non-translated Belgian Dutch.

Lapshinova-Koltunski et.al. conducted substantial work on register variation in human and machine translation. Lapshinova-Koltunski (2015) applied automated text classification methods to analyse functional variation in English-German translations. Two machine learning algorithms were tested to classify the texts into translation methods – human or machine translation respectively, and split into the following genres: political essays, fictional texts, instruction manuals, popular-scientific articles, letters of share-holders, prepared political speeches and touristic leaflets accordingly. The accuracy obtained was 60.5% for distinguishing between human-produced and machine-generated translations, and 45.4% for determining the genre of the texts. Although the methodology had yet to be refined to obtain more acute results, the study resumed that text classification approaches can be used to trace the discriminating features of genres and translation methods.

In search for the *registerness*, Lapshinova-Koltunski and Vela (2015) employed text classification methods to investigate whether translations from English into German fit into the conventional rules of the target language in terms of registers. A set of lexico-grammatical features was used to train two classifiers, k-nearest-neighbors (KNN) and support vector machines (SVM), originally authored German texts and test on comparable English-to-German translations. The experiment showed that translations can be categorised based on register characteristics, with 80% of precision was achieved for the classification of original German texts with the register features, although some of the registers were difficult to identify. In a subsequent study, Lapshinova-Koltunski (2017) continued the exploration of the correlation between the methods of translation and the registers that together with the language direction the authors refer to as *translation*

varieties. The experiment involved unsupervised machine learning techniques and statistical methods to capture variations in translation derived from these two factors. The research highlighted that register is one of the main sources to which linguistic variations in translation could be attributed.

Kunilovskaya and Corpas Pastor (2021) conducted a methodologically robust study that, apart from tracing the source or target language fingerprints, aimed at finding whether register diversity is a factor that influences English-to-Russian translations. The feature set designed for the study comprised 45 morphosyntactic features that were obtained from UD annotation, and a set of lexical indicators of translationese as discussed in previous sections. During the preliminary experiment, the full set of 56 features together with the morphosyntactic and lexical indicators separately were tested on all the text types across all the registers using the SVMs classifier. It was demonstrated that the morphosyntactic linguistic indicators yielded better results as compared to the lexical set. The evidence suggested that the similarity between the same registers in two different languages was higher than the similarities between different registers in the same language.

Another test was aimed at finding whether the classifier could accurately model the register diversity in original texts in both languages. As a result, there was a clear division between the two languages, with fiction and news commentary registers plotted closer to each other than general mass-media and popular science texts. The popular science domain showed the most prominent cross-linguistic register differences and mass-media texts demonstrated much in-category variation along the *register* axis in both languages. The findings suggested that the feature set was effective enough to accomplish the task. Based on all 56 features, the SVMs algorithm predicted each register in the two languages with around 97% accuracy. As anticipated by the authors, for this particular task the lexical features delivered better performance, as morphosyntactic indicators alone were able to obtain only 78% of accuracy for English and 81% for Russian. A major conclusion of this study is that some of the characteristic traits of translations can be explained by the shining-through effect, but others should be attributed to the established practice and professional conventions operating in translation of each register.

3 Methodology

3.1 Data Collection

Parallel and monolingual corpora that can be employed for a cross-register study of translationese are very scarce. We had to obtain components for our research corpus from diverse sources. We begin by outlining the criteria for collecting the data and then describe the chosen subcorpora.

3.1.1 Criteria for Selecting the Data

To ensure the quality of the data used in this study and to facilitate preprocessing, we have established several data selection criteria.

Availability for download

It was essential for the experiment to find publicly available datasets that could be used for research purposes locally. For instance, The CREA (Corpus de Referencia del Español Actual) [1], which is the National Corpus of the Spanish language, would have uniquely suited as the a source of originally authored Spanish texts because it embraces a variety of domains, including science, technology, fiction, politics, economy, religion, health, arts, leisure, and ordinary life. However, this corpus is only available for word search and concordance queries through its web interface.

Translation direction

For the parallel datasets, we have selected only the corpora of translations from English into Spanish and their sources. Scielo project [2], for instance, contained biomedical scientific publications and thus would have served as the primary material for a medical domain. However, the translations in the corpus were in the reverse direction, i.e. Spanish-into-English, which compelled us to discard this source.

Register diversity

Register-diversified corpora are key to our experiment as we aim at exploring the impact of register on the properties of translations (register-specific translationee). Initially, we were considering a wide range of registers from such corpora as COVID-19–HEALTH Wikipedia dataset [3], IBECS (Spanish Bibliographical Index in Health Sciences) [4], the EU bookshop [5], IULA Spanish-English Technical Corpus [6] representing biomedical, technical, and other domains. Regardless of register diversity, most of the sources lacked other important components.

Reference corpora

One of the criteria for selecting the textual material was the availability of comparable non-translations in the target language. A reference corpus should serve as a small-scale target language 'standard' for the corresponding parallel data-set and thus belong to the same or nearly the same register variety. Driven by the availability of monolingual data in the Spanish language, we eliminated the corpora for which the reference data was limited or impossible to collect. The EU book-shop and IBECS (Spanish Bibliographical Index in Health Sciences) may serve as examples.

Metadata

We have placed a special emphasis on the confidence in the translation direction and

[1] https://www.rae.es/banco-de-datos/crea
[2] https://scielo.org/es/
[3] https://opus.nlpl.eu/ELRC_2922-v1.php
[4] https://temu.bsc.es/mespen/
[5] https://opus.nlpl.eu/EUbookshop.php
[6] https://repositori.upf.edu/handle/10230/20052

the native language of the speakers that produced the texts. While searching for data, we looked for the availability of information that would clearly indicate the relations between texts in the two languages. The metadata of some sources was misleading and sometimes missing. In case of the original EuroParl corpus, the translations from English into Spanish were not marked as translations. While this subcorpus is widely used for training machine translation engines (and this is increasingly criticised), it is less appropriate for research on translationese. Therefore, we opted for the translationese-research-adapted version of EuroParl, published by (Karakanta, Vela, and Teich, 2018) and thoroughly checked the metadata for other registers included in the experiment.

Alignment level

The methodology of our experiment required that the translations and their source texts should be aligned at the document and/or sentence level. We were interested in preserving the natural text boundaries and the sequence of sentences in every document included in our research corpus. For instance, the ParaCrawl [7] corpus contains web-scraped content where the sentences and the website sections were shuffled. It was not possible to extract texts in their natural boundaries from the corpus; moreover, it was not clear whether the data represented news-wire, opinion columns, or other sections of news websites.

The next section provides details on each subcorpora that we have selected.

3.1.2 Description of the Corpus

Following the criteria for data selection outlined above, we collected a multi-corpus dataset comprised of 2 registers: parliamentary debates and fiction.

Parliamentary debates corpus consists of transcribed and revised texts of speeches by Members of the European Parliament (EP) that were produced between the years 1999 and 2017. From several available versions of the EuroParl corpus (Koehn et al., 2005; Graën, Batinic, and Volk, 2014; Calzada Pérez, Marín Cucala, and Martínez Martínez, 2006; Hajlaoui et al., 2014; Van Aggelen et al., 2017), we chose the EuroParl-UdS that was adapted specifically for translationese studies (Karakanta, Vela, and Teich, 2018). Unlike other collections, the EuroParl-UdS contains rich and reliable metadata about the speaker and the language which is crucial for our study. Besides, the authors of the collection included comparable monolingual data for each target language. For our research we have selected the following sections of EuroParl-UdS:

1. parallel sentence-aligned subcorpus of English-to-Spanish translations.

2. monolingual corpus containing original texts in Spanish

An important detail about this collection is that the texts were filtered to include only the speeches delivered by the native speakers, the MPs who come from a country

[7] https://opus.nlpl.eu/ParaCrawl.php

where the source language is an official language. For the English language, the authors agreed to take only the texts that were produced by the Members of Parliament who were originally from the United Kingdom, Ireland, or Malta. The source texts in the Spanish language (used in the reference subcorpus) included only the sentences of the representatives from Spain (Karakanta, Vela, and Teich, 2018).

Fiction corpus was extracted from the collection of out-of-copyright novels that were made publicly available by Andras Farkas [8]. The books were sentence-aligned to create parallel multilingual collections. We have selected the literary classics written in English by both American and British authors and their corresponding translations into Spanish. The majority of the literary works dates from the end of the nineteenth century, and the translations were produced mostly in the twentieth century. The non-translated (originally-authored) Spanish fiction books (fiction reference subcorpus) were extracted from the source language part of the Spanish-Russian parallel corpus included in the Russian National Corpus (RNC). The RNC contains bilingual corpora with Russian as the source or as the target language (Sitchinava et al., 2012). For the study, we have selected the novels written in the Spanish language during the ninetieth and twentieth centuries by Spanish and Latin American authors.

3.2 Data Preprocessing

Data preprocessing is another important step in any research within language technologies. Although data collection is a very time-consuming process, it is not the final stage of preparing textual data for the analysis. Sufficient attention should be paid to data preprocessing and cleaning. Noisy texts or texts containing irrelevant or missing parts, especially in a research on translationese, directly affect the outcome of a machine learning algorithm. Depending on the area of application, there are certain criteria as to how texts must be fed into a machine learning model, hence raw textual data needs to be transformed to fulfil those requirements. The ultimate goal of preprocessing is to avoid the dreaded 'garbage in, garbage out' machine learning scenario.

Data preprocessing embraces numerous methods from simple understanding the collected data by manual investigation and validation, to more sophisticated techniques such as segmentation or chunking, part-of-speech tagging, stemming, lemmatisation and parsing. *Segmentation* or *chunking* is the division of text into meaningful units such as words, phrases, paragraphs. *Part-of-speech (PoS)* tagging is the process of adding a PoS label to each token in a phrase, sentence or text. Part-of-speech tagging is complicated by the fact that the same token in different sentences may correspond to different parts of speech and therefore have a different meaning.

Stemming is the process of reducing a word to its base by discarding the endings or suffixes. Unlike stemming, *lemmatisation* uses vocabulary and morphological analysis

[8] https://farkastranslations.com/bilingual_books.php

with the goal of removing only inflectional endings and returning the base or dictionary form of a word, known as the lemma. *Parsing* is the process of mapping a linear sequence of lexemes (words, tokens) of a language in relation to its formal grammar. The result is usually a dependency or constituency tree. In the next subchapters we will discuss the methods of data preprocessing that were applied to our dataset.

3.2.1 Extracting Texts from TMX Format

There are several file formats in which large textual corpora are and distributed. Corpora in plain text format are less common. The most widespread and standard file formats for textual data are different kinds of Extensible Markup Language (XML) documents such as Standard Generalized Markup Language (SGML), Translation Memory Exchange (TMX), Text Encoding Initiative (TEI), etc. SQL database file formats such as MySQL, PostgreSQL, Oracle, and their derivatives, are also common for projects involving large textual data. The fiction subcorpus of our dataset was initially stored in TMX format. We extracted source texts and their translations from TMX format and saved them in plain text format.

In TMX files, information is presented in a segmented form, where each segment consists of a pair of sentences or phrases: the original and its translation. Text segments are usually sentences, but as the segmentation process is carried out by a computer-assisted translation tool using a set of rules based on regular expressions that take into account the sequences of certain characters, segments usually match sentences. This is because translation memories do not usually store larger unit such as paragraphs because the probability of finding the same or similar paragraphs in a text is very low; nor do they relate smaller units such as words, because the human translator does not work with these units in isolation.

An TMX document can be structurally divided into two parts. A specification of the container format, i.e. the top-level elements that provide information about the file as a whole and about the entries. In TMX, an entry consisting of aligned segments of text in two or more languages is called a translation unit, a <tu> element. A specification for the low-level meta-markup format for the content of a translation memory text segment is as follows. The root element, which encompasses the whole TMX document, is <tmx> and contains two elements:

- <header>: which contains metadata about the document;
- <body>: which contains the collection of translation units,

In turn, this element contains variants of translation units for a given language, which is defined with the xml:lang attribute, in the the <tu> elements and which may contain:

- <seg>: containing the text of the given segment;

- <note>: used to add comments;
- <prop>: allows defining properties of the parent element (the element containing <prop>).

These properties are not defined by the standard and can be used for any purpose. Therefore, to solve this formatting task we wrote a Python program that retrieved the original sentences by the tag <tuv xml:lang="en"><seg> and their corresponding translations under the tag <tuv xml:lang="es"><seg>. The originals and the translations were saved in separate files. Each literary work received a unique identification number for further processing.

3.2.2 Chunking

One of the initial processing stages in a machine learning pipeline is to split a dataset into training and test sets, usually with the ratio of 70 to 30 (or 80 to 20) respectively. The idea is to train a machine learning model on the training data, and then to check with the test data how well it has learned. The data should be divided in such a fashion that there are enough observations per each category in classification so that the machine learning algorithm could give acceptable performance. The subcorpus of political debates that we chose for the experiment was produced by Alina Karakanta (Karakanta, Vela, and Teich, 2018). It meets these requirements for a machine learning setup. However, the fiction subcorpus contained only a few long books for non-translations and translations respectively, and needed to be split into more segments before we could proceed.

Initially, we selected eight novels for the fiction subcorpus. Obviously, eight instances are not enough to train and test a machine learning model. Therefore, we decided to increase the number of observations by chunking the texts. One way of doing it could have been through dividing the texts into chunks of 300 lines each. However, with this sort of partitioning, we could have lost the correspondence between the original and the translation. We divided each novel by the chapters instead. For this we wrote a Python programme that would use a regular expression to find the beginning of each chapter and write the new chunks into separate files. During this process, we found that one of the novels translated into Spanish contained fewer chapters than the original novel in English. The translation was probably an adapted version for teenagers. As a result, we had to exclude this novel from the fiction subcorpus. This highlights the need to thoroughly manually look into the data for experiments at the preprocessing stage. Table 1 has the details on the word count, sentence count and text count for the preprocessed corpus.

After manual examination of the prepossessed corpus, we found that there were some inconsistencies in the fiction subcorpus. Mainly, it was due to periods after ‘Mr., ‘Mrs., ‘Sr., ‘Sra. and other contracted forms being misinterpreted automatically as end-of-sentence punctuation. To avoid noise at the stage of classification with machine learning

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 after chunking
(src = source, tgt = target, ref = non-translations in TL)

algorithms, we double-checked and manually corrected all the wrong line breaks in the whole fiction subcorpus before we proceeded to parsing.

3.2.3 Dependency Parsing

The extensive research on structural grammar and morphosyntactic annotation of human language yielded several techniques of automatic syntactic analysis of texts. The automatic syntactic analysis can be based on immediate constituents as proposed by Leonard Bloomfield or on the dependency relation, an approach attributed to Lucien Tesnière. Both principles arose within the framework of structuralism, and they embody a purely formal approach to syntactic categories, which makes them suitable for automatic analysis.

The purpose of the automatic syntactic analysis is to identify the basic syntactic structures and to establish syntactic relations between the words in a sentence. This is, a syntactic parser receives the input data, which is a text, and returns the same text with a mark-up of syntactic structures in the form of dependency trees. In doing so, a parser uses a model trained on an annotated corpus called a treebank. The model is trained to predict the labels (morphological categories, syntactic relations) assigned by human annotators in a treebank. The annotation of the treebanks follows a set of formal grammar rules that are learnt by the model and reused to predict labels.

There are several dependency parsers available, e.g. Stanza Dependency Parser [9], Stanford CoreNLP [10], spaCy Dependency Parser. One of the most effective and attested parser at the moment is offered by the Universal Dependency Project (UD Pipe). This is an open-community framework for morphology and syntax annotation that supports more than 100 languages and is under constant development. The project was first released in October 2014 with the aim of creating one common set of concepts or meta-language that would be used to describe syntactic relationships in natural languages. It combined

[9] https://stanfordnlp.github.io/stanza/depparse.html
[10] https://stanfordnlp.github.io/CoreNLP/index.html

elements from several initiatives developed earlier by researchers at Google and Stanford University (De Marneffe et al., 2014; Rosa et al., 2014; Tsarfaty, 2013; McDonald et al., 2013), as well as the CoNLL-X format, which was eventually updated to CoNLL-U. As the UD Project evolves, it becomes more akin to a universal grammar, aiming at effective and precise syntactic mark-up, with a special emphasis on "cross-linguistic parallelism across languages and language families"[11].

UD pipeline has been used in research on machine translation and translationese (Kunilovskaya and Kutuzov, 2017; Kunilovskaya and Corpas Pastor, 2021). It is particularly useful for translationese studies because it aims to provide a common footing for comparison of morphosyntactic properties across languages and language families despite the divergences in the word order, morphology and the function words. Besides, with the tools available within the framework of Universal Dependencies it is possible to not only define the syntactic relations between the components of a sentence, but also to annotate and extract morphological features at a word level that can be further used for experiments in translation studies.

The UD parser comprises of the following components: sentence segmentation, tokenisation, POS tagging, lemmatisation and dependency trees mark-up. In UD framework, segmentation and tokenisation are done in combination. The fundamental annotation units of UDP are the syntactic words. Importantly, these are not phonological or orthographic words. While orthographically concatenated, contractions in languages like English and Spanish are deconstructed so that the constituent syntactic components may be labelled properly. Specifically, clitic forms, for example the Spanish 'dimelo – 'say me this would be split into 'di me lo. The contracted forms are also undone, as in the English 'cannot that would be split into 'can not. Yet, the CoNLL-U format provides for the preservation of the original contracted form of the words, which is accompanied with the annotation of its components. There is a special mark-up for compound words and multiword expressions which are annotated in a way that their syntactic weight depends on the sum of their constituents.

3.3 Features Extraction

To perform machine learning experiments, the input texts must be represented in numeric form in some feature space. In other words, in the ML setting, each document (or a chunk of text) becomes an observation that exists in the form of a numerical feature vector. Our learning algorithm exploits 51 language-independent features (See Appendices) that represent the textual data in a multidimensional vector space.

The values of each component of a vector are the normalised frequencies or ratios of a particular linguistic item in the respective document. From the whole feature set, 29 attributes are based on matching the morphological and syntactic tags, and their

[11] https://universaldependencies.org/introduction.html

combinations from UD annotations. Appendix 5 has the feature names and their shorthands that are used when discussing the results and a brief explanation of each feature.

For example, *finites* are the finite forms of the verbs that were collected by matching *Verb-Form=Fin* tag in the UD-annotated documents, *ppron* stands for the frequency of personal pronouns extracted by the combination of *PRON* tag with *Person=* attribute, excluding possessive pronouns tagged *Poss=Yes*, while *nn* is the number of nouns per sentence.

This subset of features was then complemented by most common and attested textual parameters, established as effective translationese indicators in the earlier research, such as type-to-token ratio (TTR) and sentence length. With regard to TTR (the number of the vocabulary items for an individual observation [document] divided by the total word count in this observation), note that this research relies only on the content lemmas (lemmas of nouns, verbs, adjectives, adverbs) to count the number of both types and tokens. This effectively excludes the function words and allows for more comparable results across languages with different typological systems (more analytical vs. more synthetic languages).

Besides, we included other features that are the averaged frequencies of UD relations tags. Out of the total 37 dependency tags available from the UD framework, we managed to use 22 (see Appendix 5). We had to modify the list of features to exclude *cop* (copula verbs), *conj* (the connection between the two items linked by a coordinating conjunction), *csubj* (a clausal syntactic subject of a clause), *root*, *det* (a relation between the nominal head and its determiner), *punct* (any punctuation in a clause), because it is duplicated in custom features that employ list-based filtering at the extraction time. We had to omit the UD tags *clf*, a classifier that is intended for Asian languages, and *dislocated*, a tag used to indicate fronted or postposed components that do not meet the regular basic grammatical connections of a phrase, because they returned zero counts for all our observations. The values for these features are counted as the text average of the counts for a relation in each sentence. We also eliminated from the set *discourse*, *expl*, *goeswith*, *list*, *orphan*, *reparandum*, *vocative* because they produced zeros in most texts. In some experiments with Stratified GroupKFold classification, it was found that in all observations there were zeros for the 7 features mentioned above.

For feature extraction this this part, we adapted the rules developed for a similar research on English-Russian language pair and made publicly available in Kunilovskaya, Lapshinova-Koltunski, and Mitkov (2021)[12]. Generally, we proceeded from the assumption that the bigger the pool of features from which you can select, the higher the chances that we will 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

[12] https://github.com/kunilovskaya/translationese45

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

Following Kunilovskaya and Corpas Pastor (2021), the discourse markers are classified according to the descriptions in Halliday and Hasan (Halliday, Hasan, et al., 1989) and in Biber et al. (Biber, 1995) into additive, adiversative, causative, temporal sequential and epistemic. When scompiling 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 includes 46 items such as 'however', 'on the other hand', 'in spite of', etc; the comparable Spanish list has items like 'a diferencia', 'al contrario', 'ahora bien', 'a pesar de' etc. Orthography and punctuation were also used to distinguish between different patterns of discourse connectives (e.g. 'furthermore' and 'Furthermore' were matched separately). The lists of discourse markers for English and Spanish languages are included in Appendix 5.

To offer a comparison for morphosyntactic representation, detailed above, we also generated automatic alternative representations for the documents in our research corpus. To this end, we used a *tf-idf* algorithm as implemented in *scikit-learn*[13], a well-known library for machine learning in Python, and a pre-trained Sentence Transformer model from *Hugging Face* repository, called *stsb-xlm-r-multilingual*[14].

Tf-idf, or term frequency-inverse document frequency, is a statistical measure that determines the weight of each element in the generated term-document matrix. A term, or a word, is selected from each document and its weight is calculated based on the frequency of occurrence in the document and in the entire sample (i.e. its prevalence). In the TF weighting scheme, all words are of equal importance. This statistical measure decreases the weight of terms that are found too frequently and increases the weight of terms that have less frequency. Obviously, the word *european* will occur in many texts of the debates subcorpus, hence this word does not carry much information in determining the weight of a document. To tackle this disadvantage, for each word the algorithm tf-idf counts its inverse document frequency. For our experiments, we applied this algorithm on the character level and according to this created vectorised representations of each

[13] https://scikit-learn.org/stable/
[14] https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual

document in the corpus.

The pre-trained embedding model comes from a family of contextualised embedding models trained using Sentence-BERT (SBERT), a state-of-the-art framework, which uses Siamese and triplet network architectures to compute sentence representation. SBERT is an enhanced version of BERT-like models that substantially reduces the interference time of BERT. SBERT can be used a generic name for contextualised embedding models fine-tuned on various sentence-level tasks such as question-answering, sentiment, paraphrasing, etc. The model we used was fine-tuned on multi-lingual semantic textual similarity task (Reimers and Gurevych, 2019).

3.4 Experimental Setup

Our experiments are based on machine learning methods of text classification and consist in automatic categorisation of documents into predefined ontological classes (e.g. translation vs. non-translation) combined with feature selection and univariate statistic analysis. Exploring the outcomes of the automatic classification with regard to the most informative features as well as the comparison of the algorithm performance across the registers can shed light on the properties of translated texts in these registers. For our research, we considered a number of machine learning (ML) algorithms for text classification (see Section 2.2.2). However, despite a wide variety of such methods available today, not all of them are easy to implement, effective, have low computational cost in training and classification, and acceptable performance.

Several studies on translationese successfully applied the Support Vector Machines (SVM) to text classification (Baroni and Bernardini, 2006; Kurokawa, Goutte, and Isabelle, 2009; Lapshinova-Koltunski, Bizzoni, et al., 2021; Kunilovskaya and Corpas Pastor, 2021) and some of them reported higher accuracy results as compared to other machine learning methods (Ilisei et al., 2010; Van Halteren, 2008). In our study we rely on the linear SVM classifier to distinguish pairwise between three types of texts in two registers: originally authored text in the English language, their translations into Spanish and comparable Spanish texts within the same register.

First, we vectorised our corpus by extracting morphological and syntactic features listed in the previous section. One of the essential aspects of our experimental setup for SVM classifier was the standardisation of the input. The reason behind this is that in our experiments the numeric values for each linguistic feature had a lot of difference in variance, which affected the classifiers' behaviour. To cast our features into the same scale, we applied the scikit-learn *StandardScaler()* preprocessing function to our dataset. As a result, the feature values were standardized to have the mean of zero and unit variance of 1. This operation is also known as *z-transformation*.

The algorithm was then trained on these feature vectors. During training, SVM calculates the pairwise distances between the datapoints from each class, determines the closest

datapoints (called support vectors) and define the maximum hyperplane separating these support vectors. The important parameters for the SVM are the C regularisation parameter, the kernel type, and the parameters defined by the kernel. For our experiment with linear kernel, we needed only one parameter and we used the default settings in *scikit-learn* library[15], a a free software, providing a set of implementation for ML experiments in Python: C=1.0. C parameter regulates the complexity of the model, higher values of these parameters result in a more complex model: the model has higher tolerance for errors and, hence, is more overfit to the training data. In some of our classification experiments the classes are 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. It means that we repeated the training and testing cycle seven times and built seven models for our problem, each on a different split of the data into training and test sets. The number of splits corresponds to the number of novels in the smallest parallel fiction corpus. As we explained in the previous section, we partitioned each novel into multiple chunks. This corpus structure warrants the implementation of "novel-aware" folds in the first place 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, no parts of which were 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 Sentence Transformer model. The vectorisation setup in each case was as follows.

Each text was represented as a vector of tfidf-weighted frequencies of 5000 most frequent character 5-grams seen in the whole collection. We left the default settings for 'lowercase', 'stop words', 'unit norm' and 'inverse-document-frequency reweighting' parameters.

From *Sentence Transformer library* we imported the pre-trained *stsb-xlm-r-multilingual* model [16]. We then tokenised the sentences and generated an embedding for each sentence, that is each of the sentences was assigned a 768-dimensional vector. To get a text vector we averaged sentence embeddings dimension-wise (this operation is known as *mean-pooling*).

To avoid using SVM on a dataset where the ratio (number of observations / number of features) is relatively small, for high-dimensional representations we designed a simple one-layer neural model which uses *Adam optimiser* and *binary cross-entropy* as the loss

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

[16]https://public.ukp.informatik.tu-darmstadt.de/reimers/sentence-transformers/v0.2/

function. *EarlyStopping class* was set to monitor the accuracy at the validation stage. It is designed to finish the training if the testing set does not grow by 0.001 for three epochs in a row. The neural model was compiled with one dense hidden layer made of 16 nodes/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. In our case it was the class of translations. The output layer works in the following way: if the output probability is high (ex. 0.9) the doc is predicted as a translation. *Sigmoid activation function* transforms input values that are much larger than 1.0 to 1.0, values much smaller than 0.0 are snapped to 0.0.

4 Results and Discussion

In this chapter, we first present visualisations of our results based on the three types of vectorisation: morphosyntactic features, tf-idf and sentence embeddings. Over the next two subsections, we provide the results of the register classification by language and translationese classification by register, the latter being the focus of this study. In Section 4.4 we report five best translationese indicators in each register. We round off this study with the results of significance tests for each morphosyntactic feature based on frequencies in the three text types (sources, targets and references) and make conclusions about the translationese effects observed in each register.

Our main focus is on the behaviour of human-interpretable 51 morphosyntactic features: we conduct feature selection and univariate analysis based on these features (see Sections 4.4 and 4.5 below). However, the performance of an automatic classifier as well as the results of visual analysis on this feature set are thrown into perspective of the performance/results on tf-idf features and text-level-averaged sentence embeddings.

4.1 Visualisations of the Data

To obtain a general understanding of the locations of translated and non-translated registers in the feature (vector) space created by the three numeric representation (morphosyntactic features, tf-idf, sentence embeddings), we projected the data into a two-dimensional space and generated scatter plots for all data and for each register separately. Principal component analysis (PCA) was used to project a multi-dimensional vector space to a 2-dimensional space (that is each vector representing a datapoint (document or chunk in our setting) is reduced to two components in such a way that the distances between the vectorised datapoints in the new 2D space are retained as much as possible). Naturally, a 2D projection introduced distortions to the actual similarity relations between datapoints but it allows to get a general view of how well the features capture the categories in the data.

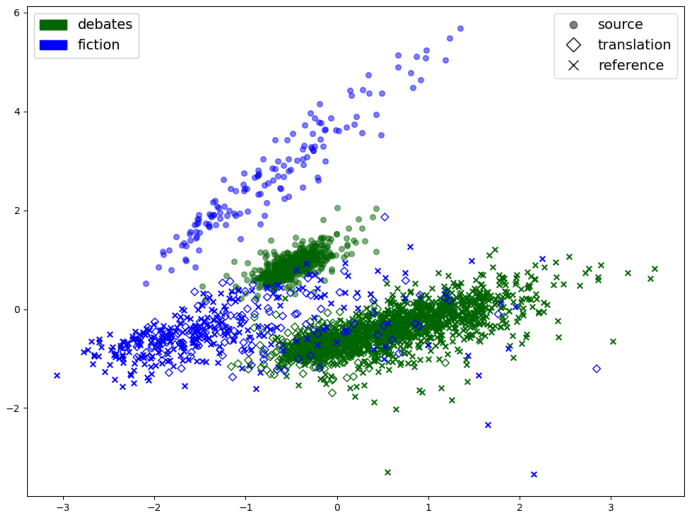


Figure 1: PCA 2D projection of six text categories (morphosyntactic features)

Figure 1 is a scatter plot that represents all six categories of texts: 2 registers and 3 text types. Each marker in the graph represents a text by only two coordinates (dimensions) obtained through PCA reduction 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 registers, even though in different directions: for fiction (blue) to the right and for debates (green) to the left. Translations seem to slightly gravitate towards 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 (see Section 4.2).

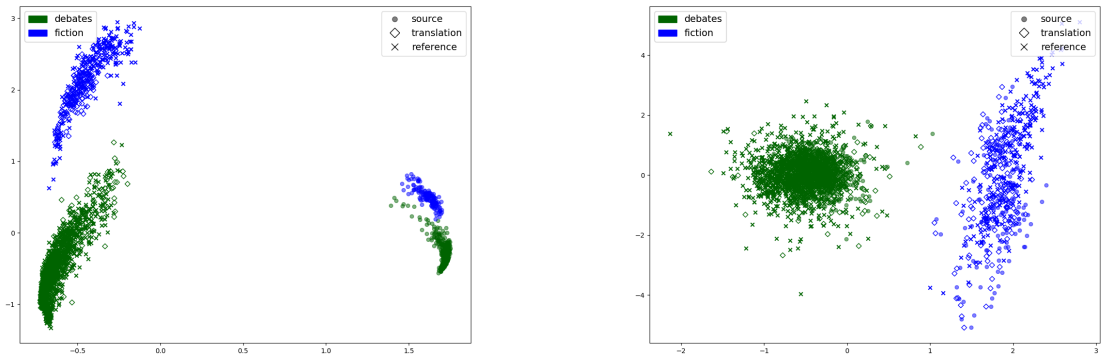
On the other hand, the areas occupied by the two registers at least in the target language seem to be more clearly defined on x-axis (see the lower two clouds of blue and green markers). At the same time the source language registers, represented by solid dots, are not that well separated horizontally, 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 vector representations, that is on tf-idf vectors and mean-pooled sentence vectors generated by a pre-trained Transformer-architecture model

(see Section 3.3).

In graph (a) (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 take the x-axis into account, the translations and reference texts in the TL of the two genres will be located on the left and the originals in the two genres on the right. Notably, the distance between languages is greater than the distance between registers, marked on the y-axis. This shows that tf-idf captured the language and register classifications well, while the text type distinction focused in this work - translations vs non-translations - seems to be very poorly captured.

Graph (b) (Figure 2) has a 2D projection of the SBERT-based embedding for our observations. The x-axis shows the difference in registers, which in the case of debates is a circle-shaped cloud, while in the case of fiction the texts are scattered in a form of an ellipsis. 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 hypothesise that the PCA reduction of the contextual embeddings oversimplifies the text distinctions that might be reflected in these vectors and captures only the strongest parameter - register.



(a) based on tf-idf vectors

(b) based on stsb_xlm_r vectors

Figure 2: PCA 2D projection on alternative representations

Figure 3 provides a different perspective on the analysis. It shows the density of the values on the first dimension of the PCA-reduced data. The x-axis has the continuum of values, while the y-axis registers 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 that was discussed in Section 2.1.1

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 4). 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

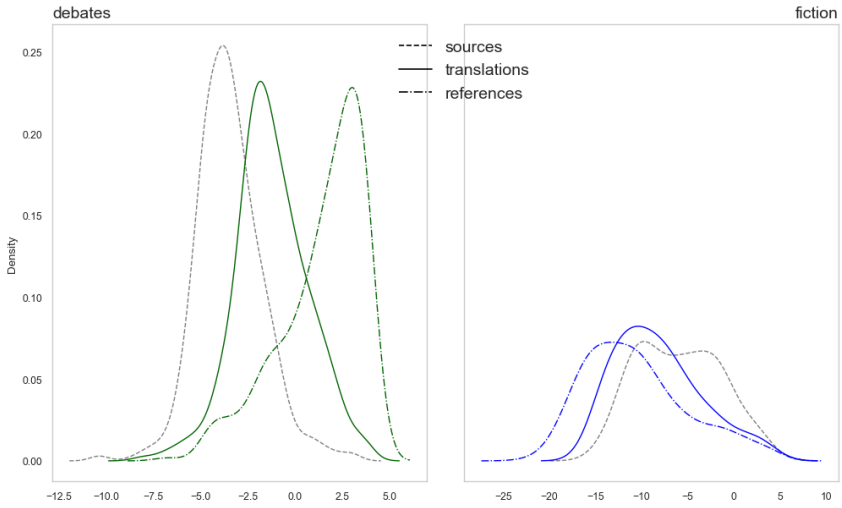


Figure 3: Distribution of values on PCA dimension 1

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.

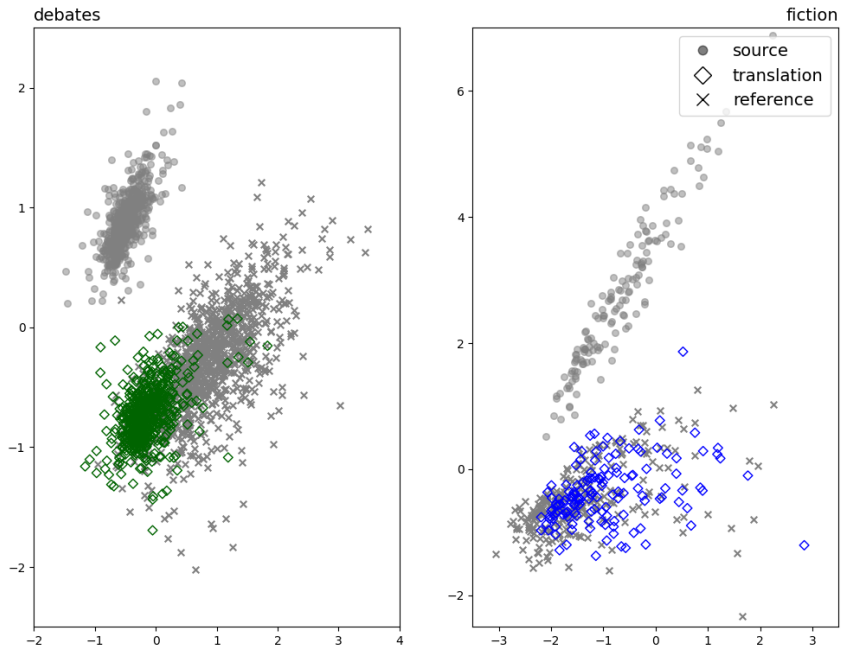


Figure 4: Sources, translations and non-translations by register

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), and translations into Spanish (Table 3). The results of three binary classifications are reported for a linear SVM classifier with default settings as described in 3.4, and for morphosyntactic features as input only.

The results for the experiment on non-translations are systematised in Table 2.

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

Table 2: Accuracy and F1-score for monolingual register classifications

For comparison, Table 3 reports the register classification results based on translations into Spanish.

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

Table 3: Accuracy and F1-score for classification on Spanish translated registers

Registers of English source texts were classified with 100% accuracy. Debates and fiction originally written in Spanish were predicted with the 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, Spanish translated registers were even more distinguishable than non-translated ones, achieving the accuracy of 99.6% with 3 misclassified instances out of 680 observations in this experiment.

These results indicate that the morphosyntactic feature capture the register distinctions very well in both languages.

Additionally, we performed feature selection using sckit-learn *SelectKBest* function, which by default runs *analysis of variance test (ANOVA)* on each feature independently (it is a univariate analysis), and returns the top N features most associated with the class label, based on the ANOVA f-score[17].

The five most register-indicative features selected for non-translated language were:

English nn, wdlength, ppron, pasttense, mhd

Spanish determ, ppron, nnargs, nn, adp

For translations into Spanish this top includes: *ppron*, *attrib*, *determ*, *simple*, *sentlength*.

Interestingly, the results of all three experiments indicate that differences between registers lie in their deictic properties, that is in the cumulative frequency of personal pronouns (and associated frequencies of nouns). Note that these are very strong predictors

[17] see https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

of register in all cases: the quality of the classification fell by only a few percent when we used these five features instead of 51. The detailed analysis of these findings are outside of the scope of this project.

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 is 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.

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

(a) linear SVM

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

(b) neural classifier

Table 4: Debates: Accuracy and F1-score for translationese classification

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

(a) linear SVM

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

(b) neural classifier

Table 5: Fiction: Accuracy and F1-score for translationese classification

Tables 4 and 5 present the results for two classifiers and three types of vectorisation on the two registers.

On the **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 performed on par with the SVM (debates: 96.6% accuracy with 96.2% F1-score; fiction: 81.1% accuracy and 78.2% F1-score).

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 with neural one (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).

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 positive: 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 representation with both classifiers archived high accuracy on the texts from the debates subcorpus, and slightly lower accuracy on the fiction subcorpus. It is in line of with the intuition: fiction in translation is expected to be more adapted to the TL norm and exhibit less influence of the ST and SL due to less emphasis on accuracy and higher aesthetic standards in this register. Translators might enjoy a higher level of freedom of expression; much more is left to their discretion than in the translation of parliamentary debates, where professional guidelines and regulations are in place.

4.4 Best Translationese Indicators by Register

One of the goals of this project was to identify the features that are most associated with either translated or non-translated text categories, and to compare them across registers. To this end, we used *analysis of variance (ANOVA)* test as described in Section 4.2. Table 6 reports the results of ANOVA-based feature selection for the two translationese classifications and the performance of a SVM classifier on them.

	top 5	acc (%)	F1 (%)
debates	sentlength, cconj, mdd, attrib, content_TTR	90.6	89.9
fiction	advers, possp, pasttense, tempseq, simple	70.2	64.1

Table 6: Best translationese indicators by register and classification results on them

The results in this experiment provide for at least two observations:

1. There are no overlapping features in the two lists.

2. The results for the top 5 translationese indicators are worse for fiction (9% from the full feature set) than for debates (6% down from the full feature set).

The first observation confirms findings by (Kunilovskaya and Corpas Pastor, 2021) for English-Russian translated registers: each register generates its own type of translationese; there are no universal translationese indicators. The second observation means that translationese properties are less expressed in fiction: translations are more statistically in line with the expected TL register norm, it is only by employing a wide range of feature it is possible to obtain relatively high classification results. Besides, the results across the folds in the experiment on fiction are much more volatile than on debates, which

signals lack of universality of these features and greater variability between instances in this register. For example, F1-scores by folds vary in the range from 37.8% to 100%.

One of the best features that worked to distinguish translations from non-translations in the fiction subcorpus was the past tense (*pasttense*). This reflects the nature of narrative prose in which the past tense prevails in order to narrate events that already happened to the protagonists of a story. Obviously, the past tense also appears in the texts of parliamentary debate, but not with the same frequency. In this register, the present tense is predominant and this could be observed in Examples 1 and 2 from the debates and fiction subcorpora respectively:

Example 1 *English*: Mr President, I should like to thank you for giving me this opportunity to address the House. I want to register my protest at the lack of facilities for me, as a Deputy within this Chamber.

Spanish: Me gustaría agradecerle la oportunidad que me brinda de dirigirme a la Cámara. Quiero formalizar una protesta por la falta de facilidades personales, en tanto que diputado de ésta Cámara.

Example 2 *English*: I had no idea that you had found my occasional retreat, still less that you were inside it, until I was within twenty paces of the door.

Spanish: Hasta llegar a veinte pasos de la puerta no tenía ni idea de que hubiera descubierto mi retiro provisional y menos aún de que estuviera dentro.

Temporal and sequential discourse markers (*tempseq*) are also frequent in fiction as it is illustrated in Example 4. They structure the narrative which develops in time by specifying the sequence of events unfolding within a story. Parliamentary speeches usually contain argumentation, descriptions or direct interaction as primary types of speech and can be characterised by brevity and conciseness as exemplified in 3. Consider the following examples:

Example 3 *English*: Madam President, may I say that I fully endorse your speech earlier in which you said that one of our jobs is to provide the public with information about the work of the institutions.

Spanish: Señora Presidenta, quisiero decir que suscribo totalmente el discurso anterior en el que dijo que una de nuestras tareas es la de informar al público sobre el trabajo de las instituciones.

Example 4 *English*: After winding along it for more than a mile, they reached their own house.

Spanish: Tras un recorrido de más de una milla, llegaron a su propia casa.

Sentence length (*sentlength*) feature contributed to detect translationese in the debates subcorpus. When we manually examined the source and target texts, we found instances of longer sentences in the Spanish language as compared to their original sources in the English language. In Example 5 the English adverb *hence* is rendered into Spanish as a four-word phrase *a eso de debe*:

Example 5 *English*: Hence this debate.

Spanish: A eso se debe este debate (That is what the debate is about).

In the next Example 6 from the debates subcorpus, the Spanish translator preferred *mantener una charla (to have a chat)* to the English verb *to speak* which has an analogy of *hablar* in the Spanish language:

Example 6

English: *I arrived here last month and spoke with the architects and builders with regard to the seating arrangements within the hemicycle.*

Spanish: *Me personé aquí el mes pasado y mantuve una charla con los arquitectos y los constructores sobre la disposición de los escaños en el hemiciclo.*

Interestingly, although the Spanish language has the analogy to the English discourse marker *firstly*, which is *primero*, the translator opted for a longer version of it - *en primer lugar*, as is shown in Example 7:

Example 7

English: I arrived here today, after being re-elected to this Parliament, to discover, firstly, that I cannot get from the '0 level of the building to the first level without getting a security card from one of the security services.

Spanish: Hoy me presento aquí, tras ser reelegido miembro de este Parlamento, descubriendo en primer lugar que no puedo acceder desde la planta baja al primer piso sin una tarjeta de seguridad concedida por los servicios de seguridad.

4.5 Translationese Effects based on Univariate Analysis

We also ran a series of univariate analyses on each feature from our set to see whether a particular linguistic occurrence could reveal shining-through effect, normalization, 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 the 1) original texts (English) 2) target language translations (Spanish) 3) reference texts in the target language (Spanish), for each morphosyntactic feature, for each register subcorpora separately.

Then, we ran a t-test (significance rest) 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. 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 because we have a lot of observations rather than 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 the language gap, a particular feature could be considered as fully adapted to the original. If neither translationese nor language gap was detected, a particular feature was considered useless. Below we report the results of the univariate analysis for the fiction and debates corpora separately.

In the fiction subcorpus, the shining-through effect, including over-shining with a possible lack of contrast between the languages, was detected on the following linguistic indicators: *sentlength*, *nn*, *mdd*, *content TTR*, *attrib*, *pasttense*, *epist*, *numcls*, *possp*, *cconj*, *copula*, *determ*, *appos*, *iobj*. As it could be observed, the number of nouns per sentence (*nn*) was one of the features that revealed shining-through effect. If we look at Example 8 from our fiction subcorpus, we can see that the following English sentence contains eight common nouns while the Spanish translation has nine common nouns in total:

> **Example 8** *English*: When Mrs. Jennings attacked her one (1) evening at the (2) park, to give the (3) name of the young (4) man who was Elinor's particular (5) favourite, which had been long a (6) matter of great (7) curiosity to her, Margaret answered by looking at her (8) sister, and saying, "I must not tell, may I, Elinor?
>
> *Spanish*: Cuando una (1) tarde, en Barton Park, la señora Jennings comenzó a asediarla para que le diera el (2) nombre del (3) joven por quien Elinor tenía especial (4) preferencia, (5) materia que desde hacía (6) tiempo carcomía su (7) curiosidad, Margaret respondió mirando a su (8) hermana y diciendo: - No debo decirlo, ¿ (9) verdad, Elinor?

And this is without even taking into account the omission of the word *park* in the Spanish translation. The adverd *long* was rendered with a nominal phrase *desde hacía tiempo (lit. since some time)* and the verbal clause in *... may I, Elionor?* was translated as a noun – *...¿verdad, Elionor? (lit. True, Elionor?)*. The nominal nature of the Spanish language 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). Although, in our corpus of fiction texts shining-through

effect is more pronounced which underlines the tendency of literary texts to be more prone to translationese.

The indicators of the shining-through effect in the debates subcorpus were as follows: *sentlength*, *wdlength*, *mhd*, *content dens*, *attrib*, *addit*, *caus*, *tempseq*, *numcls*, *simple*, *nnargs*, *ppron*, *cconj*, *sconj*, *neg*, *determ*, *ppron*, *advcl*, *advmod*, *amod*, *appos*, *case*, *cc*, *fixed*, *iobj*, *mark*, *nmod*, *nummod*, *obj*, *obl*, *parataxis*, *interrog*, *ccomp*, *xcomp*. According to the univariate analysis on the whole features set that we did on the later stages of out experiments, the number of personal pronouns grew in translations of parliamentary debates from English.

Some of the features indicated a trend to normalisation in translation, i.e. on the cases when translations from English gravitated towards the norm of the Spanish language. Such are the cases of *nn*, *finites*, *advers* and *epist* in the debates subcorpus. In the fiction subcorpus, nine tags contributed to trace adaptation: *finites*, *nnargs*, *ppron*, *aux:pass*, *ccomp*, *compound*, *fixed*, *nmod*, *xcomp*. Interestingly, *finites* worked for both subcorpora as for both registers there was a difference between the frequencies of the 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 original. In Example 9 from the fiction subcorpus, we may observe that where an adjective appears in the English language, the Spanish translator opted for an adjectival clause:

Example 9

English: Mrs. Smith has this morning exercised the privilege of riches upon a poor dependent cousin, by sending me on business to London.

Spanish: Esta mañana la señora Smith ha ejercido el privilegio de los ricos sobre un pobre primo que depende de ella (lit. *a poor cousin who depends on her*), y me ha enviado por negocios a Londres.

In the subsequent Example 10 from the debates subcorpus the translator also favoured the adaptation:

Example 10

English: They have even destroyed animals with no sign of BSE.

Spanish: Han destruido animales que no mostraban ningún síntoma de la EEB (lit. animals which did not show any symptoms of BSE).

Here, the phrase *animals with no sign of BSE* is rendered with the help of an adjectival clause containing a finite verb *animales que no mostraban ningún síntoma de la EEB* (lit. animals which did not show any symptoms of BSE). Thus, the number of verbs in translation is increased which is indicative of an adaptation to the target language norm.

The normalisation in translations of the parliamentary texts was reflected in the number of nouns per sentence (*nn*). Example 11 reveals the tendency of translators to norminalise verbal components of the original sentence in English. As it will be illustrated in Example 12, the English verb *legislate* was rendered in Spanish as *introducir legislación.* Another Example 13 from the debates subcorpus underlines the nominal nature of the Spanish language where *We look forward to the body preparing...* was rendered with two additional nouns as *Esperamos con interés la creación del organismo que deberá elaborar...(lit. We look forward to the establishment of the body that is to develop...)* while *Esperamos que el organismo elabore....* As it will be shown in Example 14, the conjunctive adverb *mercifully* was translated as a noun phrase *por suerte* while *afortunadamente* is also a valid analogy in the Spanish language

Example 11

English: We will look back in amazement at how much more progress Europe made on cutting red tape on goods and businesses than it did on people.

Spanish: Si volvemos la vista atrás nos sorprenderá comprobar que Europa ha avanzado mucho más en la reducción de trámites burocráticos para las mercancías y las empresas, que para las personas.

Example 12

English: We very much welcome the reiteration of how important it is to legislate to give legally-resident third-country nationals rights as near as possible to those of EU citizens. .

Spanish: Acogemos con gran satisfacción la reiteración de que es importante introducir legislación que conceda a los ciudadanos de terceros países que residen legalmente en la UE unos derechos que se acerquen lo más posible a aquéllos de los ciudadanos de la UE.

Example 13

English: We look forward to the body preparing the Charter of Fundamental Rights, which it would be appropriate for a representative of this Parliament to chair, producing a document that can confer direct rights on EU citizens and enable them as individuals to enforce their rights in the European Court.

Spanish: Esperamos con interés la creación del organismo que deberá elaborar la carta de derechos fundamentales y sería adecuado que estuviera presidido por un representante de este Parlamento y que presente un documento que confiera derechos directos a los ciudadanos de la UE y les permita, como individuos, ejercer sus derechos en el Tribunal Europeo.

Example 14

English: Mercifully we crashed out of exchange rate mechanism, interest rates came down and we had a happy time during the 1990s.

Spanish: Por suerte, salimos del mecanismo de tipos de cambio, los tipos de interés descendieron y tuvimos una buena época durante los años noventa.

We also grouped the features that were independent from the source and target languages. These were the cases when there was no language gap detected but there was significant difference in the averaged normalised frequencies between (i) translations and source and (ii) translations and reference texts. Among language-independent features in the fiction subcorpus were *wdlength*, *mhd*, *content dens*, *addit*, *advers*, *caus*, *tempseq*, *simple*, *intonep*, *sconj*, *neg*, *propn*, *adp*, *acl*, *expl*, *flat*, *nsubj*. It was expected that the Spanish translations would contain more negative particles, because in Spanish, unlike English, double negation occurs. Example 15 from our corpus reflects the double negation in Spanish, which also adds on the sentence length differences between the two languages as discussed in Section 4.4. 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. This is the case of a language-independent translationese.

Example 15

English: So, when it comes to the issues of liberty, the EU cannot even sort out fishing or look after our farmers. .

Spanish: Por consiguiente, cuando se trata de temas de libertad, la UE no es capaz ni siquiera de solucionar la pesca o de proteger a nuestros agricultores.

Both of the had in common as the indicators the number of prepositions and postpositions *adp*, the number of clausal modifiers of noun *acl* and the number of determiners *dep*. Apart from these features, the following indicators were language-independent in the debates subcopus: *mdd*, *content TTR*, *pasttense*, *possp*, *copula*, *aux:pass*, *compound*, *intonep*, *expl*, *flat*, *nsubj*.

The rest of the features were not translationese indicators at all. For the fiction subcorpus there were ten features that did not have any relevant statistical differences: *interrog*, *advcl*, *advmod*, *amod*, *case*, *cc*, *mark*, *nummod*, *obj*, *obl*. Interestingly, there were no useless features in the debates subcorpus, all of the indicators brought meaningful results.

Eventually, we wanted to contrast out findings against the results in Kunilovskaya and Corpas Pastor (2021). For the English-Russian language pair involved in their study, three translationese features that worked for distinguishing translation of the four registers under analysis were *sconj*, *epist*, and *parataxis* with *epist* and *sconj* representing different

translationese trends in each register. Our results show that for the English-Spanish language pair the common features among the two registers were *sentlength*, *attrib*, *determ*, *cconj*, *iobj*, *orphan* and *reparandu*. Kunilovskaya and Corpas Pastor (2021) detected 18 features indicating shining-through effect in the fiction subcorpus, which was the fewest number of features pointing to this trend as contrasted with the rest of the registers. Our study confirmed this result on the English-Spanish pair. In total, the fiction subcorpus had 15 features revealing shining-through effect while for the debates subcorpus 40 features pointed to the same trend in translation. As for the debates subcorpus, it was not possible to corroborate out results against those reported in Kunilovskaya and Corpas Pastor (2021). Yet, our results confirm that translations of the fiction register show less shining-through effect.

4.6 Limitations

Two mayor limitations to every ML experimental workflow are the lack of data and the lack of reliable data. The performance of the neural classifier could have been affected by a variety of shortcomings related to the data. One of such pitfalls is the amount of data, which was scarce especially for the fiction subcorpus. The subcorpus of debates, for instance, was represented by 1570 instances, which were the sum of all 3 types of texts: originals, their translations into Spanish and comparable reference texts in the Spanish language. This was not commensurate with the amount observations in the fiction subcorpus that totaled to 401 for all three text types. At some stages of the experiment, we needed to reduce the evaluation to 3-fold group-aware setup because the increase of folds would bring biased results. As we mentioned in the previous chapters, all the texts that comprised the debates subcorpus were prepared especially for experiments on translationese and by the same researchers. The textual data for the fiction subcorpus was prepared ad-hoc and was based on the resources available for the given language pair. For the reference subcorpus, there were only 10 novel available in open access and in *txt* format.

Following on from the issue of insufficient data, which is a common thread running through almost every study in the field of translation, our initial aim was to collect at least four different registers. By doing so, we could bring to the equation various technical and mass-media texts, and thus study more patterns of translationese in English-to-Spanish translations. Another drawback that we encountered in the middle of experiments was the low quality of data, namely, poor alignment. In the very first stages of the experiments, the classifier gave perfect accuracy, which prompted us to manually check the quality of the fiction subcorpus. It appeared that many sentences were truncated after the abbreviations Mr. and Mrs. and some of the sentences that occupied several lines in the English originals would be concatenated into one line in the Spanish translations. Therefore, we opted for the manual correction of the whole subcorpus. Yet, there was still possibility for the data

to be a bit noisy. Another caveat to the corpus of literary texts is that not all the novels that made up the reference dataset were written by authors from Spain. Some works were written by Latin American authors. Therefore, the results of the experiment for the fiction subcorpus are mixed in terms of Spanish language variants. For the future line of work, we will try to compile a larger corpus of texts, add more registers and align the representativeness of each register.

In terms of feature design, we were able to implement linguistic peculiarities on the morphological, structural and textual level but not many of them were on the lexical level. The initial idea was to go beyond lexical density and type-to-token ratio as most attested lexical features within translationese studies. Yet, this task is still a huge computational challenge. Some of the grammatical features that we wanted to generate could not be captured by the machine, such as zero subject in the Spanish language. Also, it should be noted that the *interrog* feature, or the number of interrogatives per file, was not precise because we needed to delete all the sentences that were less than three words.

The list of 5 best features for debates contained *attrib* UD tag, which is the number of adjectives and participles functioning as attributes and marked as ADJ or *VerbForm=Part* with the *amod* syntactic dependency counted per sentence. After we already got the results of the experiments we discovered that the Universal Dependences annotation marked some of the English adjectives as compound instead of *amod*. For example, the English sentence 'I am absolutely certain that it will come as a rude shock to most of the British population that internal security is an area where there is any room for involvement of the European Union. the phrase 'the European Union ran as follows:

33 the the DET DT Definite=Def|PronType=Art 35 det

34 European european PROPN NNP Number=Sing 35 compound

35 Union Union PROPN NNP Number=Sing 31 nmod SpaceAfter=No

In this example, the word *European* was marked as a *compound*. Yet, in the Spanish translation *Estoy completamente seguro que para la mayoría de la población británica le va a resultar más que chocante que en la seguridad interna haya un hueco para la intervención de la Unión Europea.* the word *Europea* was annotated as *amod*:

31 la el DET Definite=Def|Gender=Fem|Number=Sing|PronType=Art 32 det

32 Unión unión PROPN 29 nmod

33 Europea europea PROPN Gender=Fem|Number=Sing 32 amod SpaceAfter=No.

5 Conclusion

In this study, we investigated the effect of registers on translations from Spanish into English in two registers. We collected a subcorpus of literary texts and a subcorpus of parliamentary debates texts as data for the study. Each subcorpus was composed of three types of texts: (1) originally authored texts in the English language or sources, (2) their translations into Spanish or translations and (3) comparable texts of the same register in Spanish or references. Such triangulation was necessary to establish whether translated texts gravitate more toward the references, i.e. the target language norm, or the originals, thus revealing the shining-through effect in translation.

Methodologically, our experiments were based on machine learning methods of text classification. We tested the performance of a linear SVM classifier and a neural classifier on our dataset. We represented our corpus with three vectorisation methods. The first method was by extracting normalised frequencies of 51 morphological, syntactic and textual features based on the UD annotations, which were complemented by most common and attested textual parameters within translationese research. Along with these indicators, we included the frequencies of discourse markers from predefined search lists as proposed in Kunilovskaya and Corpas Pastor (2021). In order to show whether a machine can actually distinguish between translations and non-translations, we used two automatic vectorisations: the first was based on tf-idf algorithms and the second was on one of the most cutting-edge neural network technologies such as SBERT-based embeddings. We included these as alternative approaches to represent documents in our experiment. Apart from the neural extension, our research develops the approach proposed by Evert and Neumann (2017) and Kunilovskaya and Corpas Pastor (2021).

In line with our research questions, the first experiment that we conducted was that of intra-linguistic register contrast based on originally authored texts in English (sources) and non-translated texts in Spanish (references). Debates and fiction originally written in Spanish were classified with the accuracy of 98.9% and the registers in the English language were predicted with 100% accuracy. The purpose of the second series of experiments was to see whether our feature set worked well for capturing translationese across registers, which is key to our study. On the morphosyntactic features in the debates subcorpus the liner SVM classifier reached 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 showed the same results in the debates subcorpus (96.6% accuracy with 96.2% F1-score) and slightly different numbers for the fiction subcorpus (81.1% accuracy and 78.2% F1-score). Both classifiers detected translationese with almost perfect precision on texts vectorised with tf-idf. In terms of contexualised word embeddings, SVM classifier spotted translated texts with 98.4% of accuracy and 98.2% of F-score and the neural classifier achieved 97.3% accuracy with the F-score of 97.0%. The results on the fiction subcorpus are also very satisfactory: 96.6% of accuracy and 96.0%

of F-score returned by the linear SVM classifier and 94.1% with F-score of 91.6% that were rendered by the neural classifier.

One of the objective of the study was to detect the features that were the most effective in translations/non-translations classification. The analysis of variance (ANOVA) confirmed that the findings in Kunilovskaya and Corpas Pastor (2021) for English-Russian translated texts across different registers: best translationese indicators for each register are different. Another outcome of the analysis of variance is that translations of literary works gravitate more towards the TL norm. We also conducted a series of univariate analysis on the whole feature set to trace linguistic indicators that revealed shining-through or normalisation in translation. Our results on the English-Spanish pair showed that from the whole feature set 14 indicators were associated with the shining-through effect in the fiction subcorpus and 34 features pointed to the same phenomenon in the debates subcorpus. This confirms the findings in Kunilovskaya and Corpas Pastor (2021) who proved that translated literary texts translated from English into Russian were less prone to shining-through effect.

As accurate as the technical implementation of the methodology performed, our corpus was uneven with respect to registers, mostly because of the insufficient amount of data for the fiction subcorpus available publicly. Even though, the classifiers showed very satisfactory results on both registers. We hope that this research will draw more attention to register as a factor in translations, not only involving the English and Spanish pair, and as an important aspect in the training of specialists in translation and interpreting and also in the development of machine translation systems. For the future work, we are planning to continue the experiments on more registers and language pairs.

References

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

Baker, Mona (1995). "Corpora in translation studies: An overview and some suggestions for future research". In: *Target. International Journal of Translation Studies* 7.2, pp. 223–243.

— (1996). "Corpus-based translation studies: The challenges that lie ahead". In: *Benjamins Translation Library* 18, pp. 175–186.

— (2004). "A corpus-based view of similarity and difference in translation". In: *International journal of corpus linguistics* 9.2, pp. 167–193.

— (2019). "Corpus Linguistics and Translation Studies*: Implications and applications". In: *Researching Translation in the Age of Technology and Global Conflict*. Routledge, pp. 9–24.

Baroni, Marco and Silvia Bernardini (2006). "A new approach to the study of translationese: Machine-learning the difference between original and translated text". In: *Literary and Linguistic Computing* 21.3, pp. 259–274.

Bernardini, Silvia and Federico Zanettin (2004). "methodologies for the investigation of translation universals". In: *Translation Universals: Do They Exist?* 48, p. 51.

Biber, Douglas (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Biber, Douglas and Susan Conrad (2019). *Register, genre, and style*. Cambridge University Press.

Bizzoni, Yuri et al. (2020). "How human is machine translationese? comparing human and machine translations of text and speech". In: *Proceedings of the 17th International Conference on Spoken Language Translation*, pp. 280–290.

Bojanowski, Piotr et al. (2017). "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Brownlee, Jason (2017). *Deep Learning for Natural Language Processing : Develop Deep Learning Models for Natural Language in Python*. eBook, p. 414. URL: http://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes06-NMT%7B%5C_%7Dseq2seq%7B%5C_%7Dattention.pdf.

Burkov, Andriy (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov. ISBN: 9781999579500. URL: http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf.

Calzada Pérez, María, Noemí Marín Cucala, and José Manuel Martínez Martínez (2006). "ECPC: European Parliamentary Comparable and Parallel Corpora/Corpus comparables y paralelos de discursos parlamentarios europeos". In: *Procesamiento del lenguaje natural, nº 37 (sept. 2006), pp. 349-350*.

Casado-Velarde, Manuel (1978). "La transformación nominal, un rasgo de estilo de la lengua periodística". In:

Casasús Josep, Maria and Luis Núñez Ladevéze (1991). "Evolución y análisis de los géneros periodísticos" en Josep Maria Casasús y Luis Núñez Ladevéze". In: *Estilos y géneros periodísticos*.

Castagnoli, Sara (2009). "Regularities and variations in learner translations: a corpus-based study of conjunctive explicitation". In:

Chesterman, Andrew (2004). "Beyond the particular". In: *Translation universals: Do they exist* 33, p. 49.

Chowdhury, Koel Dutta, Cristina España-Bonet, and Josef van Genabith (2020). "Understanding translationese in multi-view embedding spaces". In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6056–6062.

Corpas, Gloria et al. (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: Research Papers*, pp. 75–81.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

De Marneffe, Marie-Catherine et al. (2014). "Universal Stanford dependencies: A cross-linguistic typology." In: *LREC*. Vol. 14, pp. 4585–4592.

Delaere, Isabelle, Gert De Sutter, and Koen Plevoets (2012). "Is translated language more standardized than non-translated language?: Using profile-based correspondence analysis for measuring linguistic distances between language varieties." In: *Target. International Journal of Translation Studies* 24.2, pp. 203–224.

Duff, Alan (1981). *The third language*. Pergamon Press.

Evert, Stefan and Stella Neumann (2017). "2 The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German". In: *Empirical Translation Studies*. De Gruyter Mouton, pp. 47–80.

Ferguson, Charles A (1977). "Baby talk as a simplified register". In:

Fernández Lagunilla, Marina (1999). "La lengua en la comunicación política I: El discurso del poder". In: *Madrid: Arco/Libros*.

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

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

Graën, Johannes, Dolores Batinic, and Martin Volk (2014). "Cleaning the Europarl corpus for linguistic applications". In: *KONVENS*, pp. 222–227.

Granger, Sylviane (2018). "Tracking the third code: A cross-linguistic corpus-driven approach to metadiscursive markers". In: *The corpus linguistics discourse: In honour of Wolfgang Teubert*, pp. 185–204.

Hajlaoui, Najeh et al. (2014). "Dcep-digital corpus of the european parliament". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Halliday, Michael Alexander Kirkwood, Ruqaiya Hasan, et al. (1989). "Language, context, and text: Aspects of language in a social-semiotic perspective". In:

Harris, Zellig S (1954). "Distributional structure". In: *Word* 10.2-3, pp. 146–162.

House, Juliane and Shoshana Blum-Kulka (1986). *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*. Vol. 272. Gunter Narr Verlag.

Hudson, Richard (1995). "Measuring syntactic difficulty". In: *Manuscript, University College, London*.

Ilisei (2012). "A machine learning approach to the identification of translational language: An inquiry into translationese learning models". In:

Ilisei, Iustina and Diana Inkpen (2011). "Translationese traits in Romanian newspapers: A machine learning approach". In: *International Journal of Computational Linguistics and Applications* 2.1-2, pp. 319–32.

Ilisei et al. (2010). "Identification of translationese: A machine learning approach". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 503–511.

Jing, Yingqi and Haitao Liu (2015). "Mean Hierarchical Distance Augmenting Mean Dependency Distance". In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 161–170. URL: http://ufal.mff.cuni.cz/pcedt2.0/.

Joachims, Thorsten et al. (1999). "Transductive inference for text classification using support vector machines". In: *Icml*. Vol. 99, pp. 200–209.

Jurafsky, Daniel and JH Martin (2017). *Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Hoboken: Pearson Education, Inc. URL: https://web.stanford.edu/%7B~%7Djurafsky/slp3/.

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

Kenny, Dorothy (1998). "Creatures of habit? What translators usually do with words". In: *Meta: journal des traducteurs/Meta: Translators' Journal* 43.4, pp. 515–523.

Klaudy, Kinga (1996). "Back-translation as a tool for detecting explicitation strategies in translation". In:

Koehn, Philipp et al. (2005). "Europarl: A parallel corpus for statistical machine translation". In: *MT summit*. Vol. 5. Citeseer, pp. 79–86.

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

Kunilovskaya, Maria (2017). "Linguistic tendencies in English to Russian translation: the case of connectives". In: *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"(in print)*.

Kunilovskaya, Maria and Gloria Corpas Pastor (2021). "Translationese and register variation in English-to-Russian professional translation". In: *New Perspectives on Corpus Translation Studies*. Springer Nature Singapore Pte Ltd, pp. 133–180. DOI: 10.1007/978-981-16-4918-9_6.

Kunilovskaya, Maria and Andrey Kutuzov (2017). "Universal Dependencies-based syntactic features in detecting human translation varieties". In: *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pp. 27–36.

Kunilovskaya, Maria and Ekaterina Lapshinova-Koltunski (2019). "Translationese features as indicators of quality in English-Russian human translation". In: *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pp. 47–56.

Kunilovskaya, Maria, Ekaterina Lapshinova-Koltunski, and Ruslan Mitkov (2021). "Fiction in Russian Translation : A Translationese Study". In: *RANLP-2021*, pp. 739–747. ISBN: 9789544520724.

Kurokawa, David, Cyril Goutte, and Pierre Isabelle (2009). "Automatic detection of translated text and its impact on machine translation". In: *Proceedings of Machine Translation Summit XII: Papers*.

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

Lapshinova-Koltunski, Ekaterina (2015). "Variation in translation: Evidence from corpora". In: *New directions in corpus-based translation studies* 1, p. 93.

— (2017). "7 Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method". In: *Empirical Translation Studies*. De Gruyter Mouton, pp. 207–234.

Lapshinova-Koltunski, Ekaterina, Yuri Bizzoni, et al. (2021). "Found in translation/interpreting: combining data-driven and supervised methods to analyse cross-linguistically mediated communication". In: *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pp. 82–90.

Lapshinova-Koltunski, Ekaterina and Mihaela Vela (2015). "Measuring 'registerness' in human and machine translation: A text classification approach". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*, pp. 122–131.

Laviosa, Sara (1997). "How comparable can'comparable corpora'be?" In: *Target. International Journal of Translation Studies* 9.2, pp. 289–319.

— (1998). "Core patterns of lexical use in a comparable corpus of English narrative prose". In: *Meta: journal des traducteurs/Meta: Translators' Journal* 43.4, pp. 557–570.

Laviosa-Braithwaite, Sara (1996). "The English Comparable Corpus (ECC): A resource and a methodology for the empirical study of translation". PhD thesis. University of Manchester.

Lee, David YW (2001). "Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle". In:

Mauranen, Anna and Maeve Olohan (2000). "Strange strings in translated language: A study on corpora". In: *Intercultural faultlines. Research models in translation studies I. Textual and cognitive aspects*, pp. 119–41.

McDonald, Ryan et al. (2013). "Universal dependency annotation for multilingual parsing". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 92–97.

Mikolov, Tomas et al. (2017). "Advances in pre-training distributed word representations". In: *arXiv preprint arXiv:1712.09405*.

Miller, George A (1998). *WordNet: An electronic lexical database*. MIT press.

Nadal Palazín, Juan (2008). "Verdades a medias: la nominalización deverbal en los titulares periodísticos". In: *Comunicación y sociedad* 9, pp. 175–189.

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

Olohan, Maeve (2001). "Spelling out the optionals in translation: a corpus study". In: *UCREL technical papers* 13, pp. 423–432.

Osgood, Charles Egerton, George J Suci, and Percy H Tannenbaum (1957). *The measurement of meaning*. 47. University of Illinois press.

Popescu, Marius (2011). "Studying translationese at the character level". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 634–639.

Puurtinen, Tiina (1998). "Syntax, readability and ideology in children's literature". In: *Meta: journal des traducteurs/Meta: Translators' Journal* 43.4, pp. 524–533.

— (2003). "Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature". In: *Literary and linguistic computing* 18.4, pp. 389–406.

Rayson, Paul (2008). "From key words to key semantic domains". In: *International journal of corpus linguistics* 13.4, pp. 519–549.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084*.

Rodríguez-Castro, Mónica (2011). "Translationese and punctuation: An empirical study of translated and non-translated international newspaper articles (English and Spanish)". In: *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association* 6.1, pp. 40–61.

Rosa, Rudolf et al. (2014). "HamleDT 2.0: Thirty dependency treebanks stanfordized". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 2334–2341.

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

Russell, Stuart and Russel Norvig (2020). *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken: Pearson Education, Inc, p. 2145.

Santos, Diana (1995). "On grammatical translationese". In: *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics*, pp. 59–66.

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

Shlesinger, Miriam (1989). "Extending the Theory of Translation to Interpretation: Norms as a Case in Point." In: *Target: International Journal of Translation Studies* 1.2, pp. 111–15.

Shuttleworth, Mark and Moira Cowie (1997). "Dictionary of translation studies. Manchester: St". In: *Jerome Publishing* 192, p. 193.

Sitchinava, Dmitri V et al. (2012). "Parallel corpora within the Russian National Corpus". In: *Prace Filologiczne* 63, pp. 271–278.

Straka, Milan and Jana Straková (2017). "Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99.

Teich, Elke (2003). "Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts. Berlin: Mouton de Gruyter, 2003, 276p". In: *Text, translation, computational processing* 5.

Toury, Gideon (1979). "Interlanguage and its manifestations in translation". In: *Meta: journal des traducteurs/Meta: Translators' Journal* 24.2, pp. 223–231.

— (1995). *Descriptive translation studies and beyond*. Vol. 4. J. Benjamins Amsterdam.

— (2004). "Probabilistic explanations in Translation Studies: Universals—or a challenge to the very concept?" In: *Claims, changes and challenges in translation studies*. John Benjamins, pp. 15–25.

Tsarfaty, Reut (2013). "A unified morpho-syntactic scheme of Stanford dependencies". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 578–584.

Van Aggelen, Astrid et al. (2017). "The debates of the european parliament as linked open data". In: *Semantic Web* 8.2, pp. 271–281.

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

Volansky, Vered, Noam Ordan, and Shuly Wintner (2015). "On the features of translationese". In: *Digital Scholarship in the Humanities* 30.1, pp. 98–118.

Weissbrod, Rachel (1992). "Explicitation in translations of prose-fiction from English to Hebrew as a function of norms". In:

Yang, Yiming and Xin Liu (1999). "A re-examination of text categorization methods". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49.

Zampieri, Marcos and Ekaterina Lapshinova-Koltunski (2015). "Investigating genre and method variation in translation using text classification". In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 41–50.

Zanettin, Federico (2017). "Issues in Computer-Assisted Literary Translation Studies". In: *Intralinea. Special Issue: Corpora and Literary Translation Issues in Computer-Assisted Literary Translation Studies 2017*.

Zellermayer, Michal (1990). "Shifting along the oral/literate continuum: The case of the translated text". In: *Poetics* 19.4, pp. 341–357.

Zwicky, Arnold M and Ann D Zwicky (2015). *Register as a dimension of linguistic variation*. De Gruyter.

Appendix 1

The list of features based on morphological and syntactic tags, and their combinations from the UDs annotations.

1. **sentlength**: the number of words per sentence which is normalised over all sentences in the document.

2. **wdlength**: the average word length which is counted as the number of letters in a word divided by all the words in a sentence.

3. **interrog**: the number of interrogatives per file, although this feature is not totally precise because we needed to delete all the sentences that were less than three words.

4. **nn**: the number of nouns per sentence.

5. **mhd**: the mean hierarchical distance (MHD) which is calculated as the average length of all the paths from the root of a sentence to all the nodes of this sentence along the dependency structure, as proposed in Jing and H. Liu (2015). It is also perceived as the speakers difficulty index.

6. **mdd**: the mean dependency distance (MDD), also known as comprehension difficulty, is measured as the amount of intervening words between a word and its parent within the dependency relations Hudson (1995).

7. **content dens**: the number of the content words types as divided by the number of all the tokens. The following part of speech categories are considered as content: NOUN, VERB, ADJ, ADV.

8. **content TTR**: the number of content words types as divided by the number of their tokens.

9. **finites**: the finite forms of the verbs that were collected according to the 'VerbForm=Fin tag of the UD annotation.

10. **attrib**: adjectives and participles functioning as attributes and marked as ADJ or VerbForm=Part with the 'amod syntactic dependency.

11. **pasttense**: verbs in the past tense under the 'Tense=Past UD tag.

12. **addit**: additive discourse markers that are counted as the frequency of the items from the list as normalised to the numbers of the sentences in a document. The whole list of the additive connectives is detailed in Table 3.

13. **advers**: adversative discourse markers that are counted as the frequency of the items from the list as normalised to the numbers of the sentences in a document. The whole list of the adversative connectives is detailed in (Appendix 5).

14. **caus**: causative discourse markers that are counted as the frequency of the items from the list as normalised to the numbers of the sentences in a document. The whole list of the causative connectives is detailed in (Appendix 5).

15. **tempseq**: temporal sequential discourse markers that are counted as the frequency of the items from the list as normalised to the numbers of the sentences in a document. The whole list of the temporal sequential connectives is detailed in (Appendix 5).

16. **epist**: epistemic discourse markers that are counted as the frequency of the items from the list as normalised to the numbers of the sentences in a document. The whole list of the epistemic connectives is detailed in (Appendix 5).

17. **numcls**: the number of clauses in a sentence. The following UD relations appearing in one sentence were extracted: csubj, acl:relcl, advcl, acl, xcomp, parataxis.

18. **simple**: the number of simple sentences, i.e. in which there was not a word that was marked with the following UD relations: csubj, acl:relcl, advcl, acl, xcomp, parataxis

19. **nnargs**: nouns and proper names as the main verbal arguments, counted as the number of nouns and propers names in the nsubj, obj, iobj UD relations divided by the total number of these relations.

20. **ppron**: the number of personal pronouns extracted by the 'Person UD tag, excluding the possessive pronouns under the UD tag 'Poss=Yes.

21. **possp**: the number of possessive pronouns extracted by the 'Poss=Yes UD tag.

22. **intonep**: the number of indefinite pronouns

23. **conj**: the number of coordinating conjunctions

24. **sconj**: the number of subordinating conjunctions

25. **neg**: the number of negative particles

26. **copula**: the number of copula verbs extracted by the 'cop UD tag for the verb 'be in the English language and the verbs 'ser, 'estar for the Spanish language.

27. **determ**: the number of the determiners

28. **propn**: the number of personal pronouns

29. **adp**: the number of prepositions and adpositions per sentence.

Appendix 2

The list of features based on syntactic relations from the UD annotations.

1. **acl**: clausal modifier of noun;
2. **advcl**: adverbial clause modifier;
3. **advmod**: adverbial modifier;
4. **amod**: adjectival modifier;
5. **appos**: appositional modifier;
6. **aux:pass**: passive auxiliary;
7. **case**: case marking;
8. **cc**: coordinating conjunction;
9. **ccomp**: clausal complement;
10. **compound**: compound;
11. **dep**: determiner;
12. **fixed**: fixed multiword expression;
13. **flat**: flat multiword expression;
14. **iobj**: indirect object;
15. **mark**: a verbs that marks a clause as subordinate to another clause;
16. **nmod**: nominal modifier
17. **nsubj**: nominal subject
18. **nummod**: numeric modifier
19. **obj**: object
20. **obl**: oblique nominal
21. **parataxis**: parataxis, marks a relation between a word and other elements in a sentence
22. **xcomp**: open clausal complement

Appendix 3

The lists of discourse markers for English and Spanish languages

	English	Spanish
Additive	53	75
Adversative	46	40
Causative	42	58
Temporal sequential	110	63
Epistemic	64	38

Additive DMs for English:

additionally, add to this, also, alternatively, and another thing, as evidence of, as for, besides, by the same token, by the way, concerning, equally, for another thing, for example, for instance, for one thing, furthermore, i mean, in addition, incidentally, in just this way, in other words, in particular, in this regard, in this respect, in that regard, in the same way, in this regard, in this way, likewise, more accurately, more importantly, moreover, more precisely, more to the point, namely, not only, on top of that, on top of this, or else, put it another way, separately, similarly, specifically, such as, that is, that is to say, too, to put it another way, what is more, with regard to, i.e.

Additive DMs for Spanish:

a decir verdad, a saber, a su vez, a todo esto, acerca de, además, al mismo tiempo, análogamente, aparte, así, así como, así por ejemplo, así pues, así tenemos, asimismo, como muestra, con referencia a, con relación a, con respecto a, conciernemente a, concretamente, de forma similar, de hecho, de igual importancia, de igual manera, de igual modo, de la misma forma, de la misma forma, de la misma manera, de nuevo, del mismo modo, dicho sea de paso, en concreto, en cuanto a, en el fondo, en igual forma, en las mismas circunstancias, en lo que concierne a, en lo que respecta, en lo que toca a, en lo tocante, en particular, en realidad, en relación con, encima, entonces, entonces eso, entre paréntesis, es decir, es más, es que, esto es, hablando de, hay que mencionar, igualmente, más, más aún, o sea, otro aspecto, otro punto es, pongamos por caso, por añadidura, por cierto, por ejemplo, por lo que se refiere a, por lo que se refiere de, por otro lado, por si fuera poco, propósito, referente a, respecto a, sobre, sobre todo, verbigracia, y.

Adversative DMs for English:

actually, all the same, alternatively, although, anyhow, but anyway, by contrast, apart from this, as against that, as a matter of fact, aside from this, as opposed to, conversely, despite, except, however, in comparison, in contrast, in fact, in another event, in other events, in point of fact, in spite of, instead, it may be the case that, nevertheless, nonetheless, notwithstanding, on the contrary, on the one hand, on the other hand, otherwise, rather, rather than, regardless, even so, still, such not being the case, that not being the case, that said, the alternative is, though, to tell the truth, unlike, whereas, yet.

Adversative DMs for Spanish:

a diferencia, a pesar de, ahora bien, al contrario, antes al contrario, antes bien, aparte de eso, aparte de esto, así y todo, aun así, aun con eso y con todo, aun con todo, aunque, con esto y todo, con todo, con todo y con eso, de cualquier manera, de todas formas, de todas maneras, de todos modos, después de todo, empero, en cambio, en cualquier caso, en efecto, en realidad, eso sí, esto sí, excepto, incluso, más bien, no obstante, pero, pero de todos modos, por contra, por el contrario, puede ser que, sin embargo, todo lo contrario, lo anterior no quiere decir que.

Causative DMs for English:

accordingly, after all, arising from this, arising out of this, as a consequence, as a result, because, can be concluded, consequently, correspondingly, due to, following from this, for all that, for all this, for that reason, for this purpose, for this reason, from this it appears that, hence, in consequence, in such a case, in such an event, in that case, it follows from this, it follows that, on account of this, on this basis, so, so that, that being the case, the consequence of that is, the reason was that, therefore, to that end, to this end, under the circumstances, under these circumstances, under those circumstances, we may conclude that, with this in mind, with this intention, with this in view.

Causative DMs for Spanish:

así que, por esta razón, de modo que, por consiguiente, por ende, ello se debe a, por lo tanto, porque, pues, dado que, a causa de, por el hecho de que, ya que, puesto que, en consecuencia, como consecuencia, como resultado, a raíz de, por este motivo, por eso, por esto, es que, y es que, a la vista de esto, en vista de, por causa de, así pues, así, así que, con lo que, con lo cual, consecuentemente, consiguientemente, dadas estas circunstancias, dadas esas circunstancias, dadas tales circunstancias, de ahí que, de aquí que, debido a esto, debido a ello, de modo que, de manera que, de suerte que, en consecuencia, entonces, gracias a esto, gracias a ello, o sea que, o sea, por consiguiente, por culpa de esto, por culpa de eso, por este motivo, por ese motivo, por esto, por ello, por tanto y pues, por lo tanto y pues.

Epistemic DMs for English:

all things considered, apparently, arguably, as a general rule, as expected, as far as we know, as for me, assuming that, as to me, best of my knowledge, beyond doubt, can be argued, certainly, clearly, decidedly, definitely, evidently, for sure, from my point of view, goes without saying, indeed, in essence, in my eyes, in my opinion, in my view, in reality, maybe, naturally, no doubt, obviously, of course, on balance, perhaps, possibly, predictably, presumably, probably, provisionally, really, seemingly, sure, sure enough, surely, tentatively, to my mind, undoubtedly, no doubt, no doubts, looks like, it appears that, it seems, appear to, appears to, in our view, from our perspective, at any rate, in many cases, in any case, in any event, in either case, at least, either way, whichever.

Epistemic DMs for Spanish:

dicho con otras palabras, dicho en otros términos, dicho de otra forma, dicho de otra manera, de otro modo, más claramente, más llanamente, hablando en plata, más

bien, mejor dicho, por mejor decir, a ver, digamos, entonces, entonces eso, es verdad que, es cierto que, dicho esto, pues bien, deseo subrayar que, indiscutiblemente, lo más importante, lo peor del caso, por supuesto que, precisamente, a lo mejor, admitamos por el momento, consideremos, esta hipótesis, Es posible que, Es probable que, Parto de la siguiente hipótesis, Planteo como hipótesis, Posiblemente, con esto quiero decir, como se ha dicho, en otras palabras, Todo esto parece confirmar, Me gustaría dejar claro.

Temporal sequential DMs for English:

after, after a time, after that, afterwards, all in all, all this time, and then, Anyway, as a final point, as long as, as soon as, at first, at last, at once, at that time, at the same time, at this moment, at this point, basically, before that, before then, briefly, broadly speaking, by and large, by this time, concurrently, earlier, eventually, finally, first, first and foremost, firstly, first of all, five minutes earlier, five minutes later, formerly, from now on, Further, henceforward, heretofore, hereunder, hirherto, i have already noted, immediately, in a nutshell, in a word, in conclusion, in future, in general, in short, in sum, in summary, in the end, in the first place, in the future, in the meantime, in the second place, just before, generally, lastly, later, meanwhile, Next, next day, next moment, next time, Now, on a different note, on another occasion, on a previous occasion, once again, on the whole, on this occasion, over the next, presently, previously, second, secondly, simultaneously, since, So anyway, so far, some time earlier, soon, subsequently, the last time, Then, the previous moment, thereafter, thereupon, Third, thirdly, this time, thus, to begin with, to cap it all, to conclude with, to get back to the point, to put it briefly, to resume, to start with, to summarise, to summarize, to sum up, ultimately, until then, up till that time, up to now, up to this point, while, whilst.

Temporal sequential DMs for Spanish:

a continuación, a fin de cuentas, a la vez, a medida que, actualmente, ahora, al fin y al cabo, al final, al principio, ante todo, antes, antes de nada, antes que nada, con todo, de otra parte, de una parte, desde entonces, después, después de todo, en conclusión, en definitiva, en esta época, en este momento, en fin, en la actualidad, en otra época, en primer lugar, en primer término, en resumen, en resumidas cuentas, en segundo lugar, en segundo término, en síntesis, en suma, en tercer lugar, en tercer término, en último lugar, en una palabra, entonces, entre tanto, finalmente, hasta ahora, luego, más adelante, más tarde, mientras tanto, para empezar, para terminar, por ahora, por otra parte, por otro lado, por último, Por último, y como punto no menos importante, por un lado, por una parte, posteriormente, primeramente, primero, pronto, resumiendo, seguidamente, siguiente, simultáneamente, total, tras.

Appendix 4

Results of the Univariate Analysis

The results for each feature in the fiction subcorpus consisting of 545 instances are as follows:

Fiction

1. **sentlength**: EN 24.449; TGT 22.006; REF 19.593, Effect size -0.33.
2. **wdlength**: EN 4.238; TGT 5.075; REF 4.798, Effect size -0.48.
3. **interrog**: EN 0.096; TGT 0.099; REF 0.096, Effect size -0.05.
4. **nn**: EN 0.137; TGT 0.156; REF 0.164, Effect size 0.42.
5. **mhd**: EN 3.114; TGT 3.587; REF 3.252, Effect size -0.48.
6. **mdd**: EN 1.448; TGT 1.263; REF 1.119, Effect size -0.46.
7. **content density**: EN 0.36; TGT 0.386; REF 0.373, Effect size -0.72.
8. **content TTR**: EN 0.981; TGT 0.982; REF 0.977, Effect size -0.40.
9. **finites**: EN 0.981; TGT 0.982; REF 0.977, Effect size -0.40.
10. **attrib**: EN 0.853; TGT 0.816; REF 0.706, Effect size -0.25.
11. **pasttense**: EN 2.012; TGT 0.887; REF 0.596, Effect size -0.96.
12. **addit**: EN 0.03; TGT 0.637; REF 0.51, Effect size -0.55.
13. **advers**: EN 0.099; TGT 0.166; REF 0.085, Effect size -1.38.
14. **caus**: EN 0.031; TGT 0.127; REF 0.093, Effect size -0.59.
15. **tempseq**: EN 0.129; TGT 0.15; REF 0.102, Effect size -0.79.
16. **epist**: EN 0.098; TGT 0.029; REF 0.022, Effect size -0.41.
17. **numcls**: EN 1.421; TGT 1.472; REF 1.208, Effect size -0.57.
18. **simple**: EN 0.393; TGT 0.364; REF 0.446, Effect size 0.79.
19. **nnargs**: EN 0.389; TGT 0.526; REF 0.539, Effect size 0.16.
20. **ppron**: EN 0.08; TGT 0.06; REF 0.058, Effect size -0.16.
21. **possp**: EN 0.007; TGT 0.002; REF 0.015, Effect size -1.05.
22. **intonep**: EN 0.003; TGT 0.017; REF 0.014, Effect size -0.69.

23. **cconj**: EN 0.877; TGT 0.812; REF 0.693, Effect size -0.27.
24. **sconj**: EN 0.408; TGT 0.804; REF 0.648, Effect size -0.43.
25. **neg**: EN 0.296; TGT 0.34; REF 0.285, Effect size -0.40.
26. **copula**: EN 0.51; TGT 0.292; REF 0.222, Effect size -0.65.
27. **determ**: EN 0.07; TGT 0.071; REF 0.082, Effect size 0.77.
28. **propn**: EN 0.54; TGT 0.838; REF 0.668, Effect size -0.45.
29. **adp**: EN 0.09; TGT 0.126; REF 0.118, Effect size -0.41.
30. **acl**: EN 0.002; TGT 0.001; REF 0.001, Effect size 0.20.
31. **advcl**: EN 0.004; TGT 0.005; REF 0.006, Effect size 0.05.
32. **advmod**: EN 0.013; TGT 0.009; REF 0.01, Effect size 0.08.
33. **amod**: EN 0.007; TGT 0.006; REF 0.008, Effect size 0.12.
34. **appos**: EN 0.001; TGT 0.002; REF 0.003, Effect size 0.26.
35. **aux:pass**: EN 0.002; TGT 0.0; REF 0.0, Effect size XXX.
36. **case**: EN 0.02; TGT 0.019; REF 0.026, Effect size 0.13.
37. **cc**: EN 0.007; TGT 0.006; REF 0.009, Effect size 0.13.
38. **ccomp**: EN 0.003; TGT 0.001; REF 0.001, Effect size XXX.
39. **compound**: EN 0.002; TGT 0.0; REF 0.0, Effect size 0.14.
40. **dep**: EN 0.0; TGT 0.0; REF 0.0, Effect size 0.34.
41. **fixed**: EN 0.0; TGT 0.001; REF 0.001, Effect size 0.51.
42. **flat**: EN 0.001; TGT 0.001; REF 0.001, Effect size 0.43.
43. **iobj**: EN 0.0; TGT 0.005; REF 0.006, Effect size 0.19.
44. **mark**: EN 0.008; TGT 0.008; REF 0.01, Effect size 0.07.
45. **nmod**: EN 0.007; TGT 0.008; REF 0.012, Effect size 0.14.
46. **nsubj**: EN 0.016; TGT 0.006; REF 0.008, Effect size 0.20.
47. **nummod**: EN 0.001; TGT 0.001; REF 0.001, Effect size 0.12.
48. **obj**: EN 0.01; TGT 0.008; REF 0.01, Effect size 0.14.

49. **obl**: EN 0.01; TGT 0.009; REF 0.011, Effect size 0.10.

50. **parataxis**: EN 0.001; TGT 0.002; REF 0.002, Effect size 0.22.

51. **xcomp**: EN 0.003; TGT 0.002; REF 0.002, Effect size -0.06.

The results for each feature in the debates subcorpus consisting of 2109 observations are as follows:

Debates

1. **sentlength**: EN 25.449; TGT 27.842; REF 35.023, Effect size 1.79.

2. **wdlength**: EN 4.749; TGT 5.088; REF 5.219, Effect size 0.56.

3. **interrog**: EN 0.037; TGT 0.037; REF 0.023, Effect size -0.49.

4. **nn**: EN 0.188; TGT 0.196; REF 0.197, Effect size 0.06.

5. **mhd**: EN 3.627; TGT 4.034; REF 4.334, Effect size 1.14.

6. **mdd**: EN 1.571; TGT 1.407; REF 1.542, Effect size 1.29.

7. **content density**: EN 0.399; TGT 0.393; REF 0.378, Effect size -1.07.

8. **content TTR**: EN 0.967; TGT 0.969; REF 0.955, Effect size -1.27.

9. **finites**: EN 0.8; TGT 0.837; REF 0.833, Effect size -0.06.

10. **attrib**: EN 1.424; TGT 1.535; REF 1.991, Effect size 1.28.

11. **pasttense**: EN 0.884; TGT 0.456; REF 0.489, Effect size 0.25.

12. **addit**: EN 0.117; TGT 0.825; REF 1.008, Effect size 1.19.

13. **advers**: EN 0.097; TGT 0.147; REF 0.145, Effect size -0.04.

14. **caus**: EN 0.075; TGT 0.145; REF 0.219, Effect size 1.10.

15. **tempseq**: EN 0.091; TGT 0.124; REF 0.158, Effect size 0.57.

16. **epist**: EN 0.113; TGT 0.015; REF 0.017, Effect size 0.06.

17. **numcls**: EN 1.503; TGT 1.669; REF 2.006, Effect size 1.09.

18. **simple**: EN 0.281; TGT 0.246; REF 0.218, Effect size -0.47.

19. **nnargs**: EN 0.536; TGT 0.707; REF 0.711, Effect size 0.09.

20. **ppron**: EN 0.044; TGT 0.027; REF 0.025, Effect size -0.34.

21. **possp**: EN 0.004; TGT 0.011; REF 0.009, Effect size -0.81.

22. **intonep**: EN 0.001; TGT 0.012; REF 0.01, Effect size -0.81.
23. **cconj**: EN 0.803; TGT 0.828; REF 1.203, Effect size 1.73.
24. **sconj**: EN 0.492; TGT 1.058; REF 1.309, Effect size 0.94.
25. **neg**: EN 0.236; TGT 0.281; REF 0.321, Effect size 0.44.
26. **copula**: EN 0.501; TGT 0.369; REF 0.453, Effect size 0.79.
27. **determ**: EN 0.096; TGT 0.117; REF 0.118, Effect size 0.10.
28. **propn**: EN 0.623; TGT 1.08; REF 1.348, Effect size XXX.
29. **adp**: EN 0.107; TGT 0.156; REF 0.155, Effect size -0.13.
30. **acl**: EN 0.002; TGT 0.001; REF 0.002, Effect size 0.49.
31. **advcl**: EN 0.002; TGT 0.005; REF 0.01, Effect size 0.38.
32. **advmod**: EN 0.007; TGT 0.008; REF 0.017, Effect size 0.47.
33. **amod**: EN 0.008; TGT 0.014; REF 0,029, Effect size 0.49.
34. **appos**: EN 0.001; TGT 0.002; REF 0.006, Effect size 0.57.
35. **aux:pass**: EN 0.001; TGT 0.0; REF 0.001, Effect size 0.29.
36. **case**: EN 0.016; TGT 0.032; REF 0.066, Effect size 0.50.
37. **cc**: EN 0.004; TGT 0.007; REF 0.016, Effect size 0.49.
38. **ccomp**: EN 0.002; TGT 0.002; REF 0.004, Effect size 0.35.
39. **compound**: EN 0.005; TGT 0.0; REF 0.0, Effect size 0.19.
40. **dep**: EN 0.0; TGT 0.0; REF 0.0, Effect size 0.14.
41. **fixed**: EN 0.0; TGT 0.002; REF 0.004, Effect size 0.45.
42. **flat**: EN 0.001; TGT 0.0; REF 0.001, Effect size 0.31.
43. **iobj**: EN 0.0; TGT 0.004; REF 0.008, Effect size 0.40.
44. **mark**: EN 0.007; TGT 0,012; REF 0.023, Effect size 0.42.
45. **nmod**: EN 0.008; TGT 0.018; REF 0.038, Effect size 0.51.
46. **nsubj**: EN 0.01; TGT 0.008; REF 0.014, Effect size 0.42.
47. **nummod**: EN 0.001; TGT 0.002; REF 0.003, Effect size 0.43.

48. **obj**: EN 0.007; TGT 0.009; REF 0.018, Effect size 0.45.

49. **obl**: EN 0.007; TGT 0.011; REF 0.023, Effect size 0.48.

50. **parataxis**: EN 0.0; TGT 0.001; REF 0.002, Effect size 0.39.

51. **xcomp**: EN 0.002; TGT 0.002; REF 0.003, Effect size 0.39.

List of Figures

List of Tables