《翻譯季刊》

Translation Quarterly

二零二一年九月　第一百零一期

No. 101, September 2021

 ISSN 1027-8559-101

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

The Hong Kong Translation Society has entered into an electronic licensing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The full text of Translation Quarterly can be found on EBSCO Publishing's databases.

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

翻譯季刊

Translation Quarterly

香港翻譯學會

The Hong Kong Translation Society

Editors' Note

In less than three decades since Mona Baker put forward the idea of studying translation using corpora in 1993, corpus-based translation studies (CBTS) has now firmly established itself as a major research area within translation studies. It is no exaggeration to claim that CBTS has greatly strengthened translation studies as an independent discipline by conceiving translation as a legitimate and creative activity rather than a derivative one.

CBTS has continued to garner interest among translation scholars with its improved methods and techniques. In this special issue titled "Recent Advances in Corpus-based Translation Studies", we bring together various researchers working with diverse language backgrounds to share their unique perspectives and findings. We hope that this special issue can yield some new insights into corpus-based approaches to translation studies.

For a detailed introduction of this special issue, readers are referred to the first chapter "Corpora: A Lens into Translation Phenomena" written by us.

September 2021

Sara Laviosa and Kanglong Liu

CONTENTS

Corpora: A Lens into Translation Phenomena

Sara Laviosa[1] *Kanglong Liu*[2]

Address: [1]Department of Humanistic Research and Innovation, University of Bari 'Aldo Moro', Bari, Italy;
[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China
E-mail: [1]sara.laviosa@uniba.it; [2]kl.liu@polyu.edu.hk
Correspondence: Sara Laviosa
Citation: Laviosa, Sara, and Kanglong Liu. 2021. "Corpora: A Lens into Translation Phenomena." *Translation Quarterly* 101: 1-4.

In the late 1990s, when the corpus-based approach to the study of the product and process of translation and interpreting was still pioneering work, Maria Tymoczko (1998) predicted that it would play a leading role in the discipline of Translation Studies in the following decades. One of the main reasons for imagining such scenario was the capacity of corpus studies to "change in a qualitative and quantitative way both the content and the methods of the discipline of Translation Studies, in a way that fits with the modes of the information age" (Tymoczko 1998, 652). Hence, as Tymoczko contended, "Corpus translation studies is central to the way that Translation Studies as a discipline will remain vital and move forward" (Tymoczko 1998, 652).

Indeed, Tymoczko's predictions have all come true. Since its advent, Corpus-based Translation Studies (henceforth CBTS) has amply demonstrated that one of its strengths is the capacity of continually adapting modern technologies to the discipline's ever-changing needs and purposes in line with socio-cultural changes and the demands of the job market. What are the emergent exigencies of Translation Studies that CBTS sets out to address today? What are the new goals of Translation Studies that CBTS intends to pursue? The papers commissioned for this Special Issue of *Translation Quarterly* aptly illustrate how CBTS is exploring new avenues of enquiry in order to meet the current demands of Translation Studies and contribute to the achievement of its current objectives. CBTS is capable of doing so thanks to the design and availability of new corpora in different languages as well as the use of sophisticated methods of statistical analysis.

The papers we have selected for this Special Issue of the journal constitute a representative sample of recent trends in CBTS, that arise from the needs of our globalized world and

an increasingly technologized language industry worldwide. These trends pertain to two main research domains, namely descriptive and applied studies. Within descriptive CBTS, we can identify the quest for translation universals as a line of enquiry that builds on the theoretical and empirical work of previous scholars, continually refines the methodology used, and discovers patterns of translational language that enhance our understanding of the nature of translation as a distinctive form of cross-linguistic and cross-cultural mediation that is influenced by the stylistic norms established for particular genres, registers or language varieties. Indeed, as Kirsten Malmkjær (2008) suggests, general features of translational behaviour such as simplification, explicitation and normalization would be better accounted for by the norm concept and explained on socio-cultural grounds.

Within applied CBTS, research into machine translation, as part of translation aids, goes hand in hand with research into translation quality assessment, as part of translator training. Translator education in turn comprises the subdomains of teaching methods, testing techniques and curriculum planning. These ambits of scholarly enquiry aim to respond to the challenges posed by the growing impact of technology in the provision of translation services, particularly in the last decade. Another research endeavour that is making inroads into applied translation studies, especially in translation pedagogy, arises from the need to develop the ability to translate into language B. This concern is well motivated by the continuing expansion of English as lingua franca.

The first paper by **Sara Laviosa and Kanglong Liu**, "The Pervasiveness of Corpora in Translation Studies", provides the general background to the four contributions that follow on. The first author gives an overview of corpus use in Translation Studies from the early 1990s to the first decade of the new millennium. The second author assesses the state of the art of corpus studies of translation over the last ten years. In the concluding section, they pull together the main threads they laid out in their overview of the field, and then point to future directions.

In the second paper, "Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-Translated French", **De Clercq et al.** analyse the linguistic features that distinguish original French newspaper texts from English to French machine-translated newspaper texts, using four machine translation systems, that is one statistical and three neural systems. The aim is to evaluate the output of different types of MT systems for both translator training and professional purposes.

In the third paper, "A Corpus-based Approach to Profiling Translation Quality: Measuring and Visualizing Acceptability of Student Translation", **Yanmeng Liu** addresses the thorny issue of assessing the quality of student translations in a valid and reliable way. To achieve this goal, Liu develops a corpus-based approach to profiling and assessing translation acceptability, conceived as a dimension of translation quality. The approach involves the statistical comparison of lexical, syntactic and grammatical features of English texts translated from Chinese vis-à-vis original English texts. More specifically, Liu uses two corpora, i.e. the

Parallel Corpus of Chinese EFL Learners (PACCEL) and the Lancaster-Oslo/Bergen Corpus (LOB), together with machine learning methods and analytical statistics to verify assessment efficiency and display the results.

In the fourth paper, "Translation Universals in Legal Translation: A Corpus-based Study of Explicitation and Simplification", **Francesca Luisa Seracini** combines quantitative and qualitative analyses of patterns of language use in translated EU texts in Italian versus original English EU texts. These typical features of translational language are considered to be manifestations of the posited translation universals of simplification and explicitation. Seracini argues that the occurrence of these linguistic patterns provides not only evidence for the presumed existence of distinctive features of translational behaviour regardless of the influence of the source language, but also evidence for the influence of the stylistic norms recommended for given subject-specific domains in the target language and culture, such as legal language.

In the fifth and last paper, "English for a Global Readership: Implications for the L2 Translation Classroom", **Dominic Stewart** tackles the complexity of establishing what are appropriate or inappropriate translation renderings when the target language is English as a Lingua Franca (ELF). Based on examples drawn from his Italian-to-English translation classes, where different corpus-based resources are regularly used for translating tourist text aimed at an international readership, Stewart demonstrates the conundrum that teachers and students face when different reference materials (such as monolingual dictionaries for advanced learners of English) and different corpus-based resources (such as English reference and web-derived corpora) provide inconsistent evidence as to the accuracy and fluency of potential translation equivalents.

This special thematic issue "Recent Trends in Corpus-based Translation Studies" fittingly closes with a review by Cui Xu of an edited volume, *CTS Spring-cleaning: A Critical Reflection*, edited by María Calzada Pérez and Sara Laviosa. Published in a Special Issue of the journal *MonTI* in 2021, the articles indeed highlight the recent trends in the field of CBTS by placing a special focus on overlooked areas such as subtitling, travel journalism, localization, and operatic audio description. As stated by Xu in the book review, the studies have offered refreshing perspectives in the field of corpus-based translation and interpreting studies.

To conclude, based on the evidence provided by the papers collected here, we can confidently affirm that CBTS is proving to be exactly what Tymoczko (1998, 652) envisaged it would be. CBTS is promoting "the construction of information fields that suit a new international, multicultural intellectualism, providing for the inclusion of data from small and large populations, from minority as well as majority languages and cultures". CBTS is stimulating collaborative research endeavours "unimpeded by time or space". Like large databases in science, corpora make it possible to build upon past research and become "a legacy of the present to the future, enabling future research to build upon that of the present". CBTS has marked "a turn away from prescriptive approaches to translation toward descriptive approaches". Fi-

nally, CBTS is reengaging “the theoretical and pragmatic branches of Translation Studies, branches which over and over again tend to disassociate, developing slippage and even gulfs” (Tymoczko 1998, 658). Indeed, the technical and descriptive investigations undertaken within corpus-based studies have now, more than they have ever had, “practical potential and immediate applicability, not only for the teaching of translation but for the work of the practising translator as well” (Tymoczko 1998, 658).

References

Malmkjær, Kirsten. 2008. “Norms and Nature in Translation Studies.” In *Incorporating Corpora. The Linguist and the Translator*, ed. by Gunilla Anderman, and Margaret Rogers, 49-59. Clevedon: Multilingual Matters.

Tymoczko, Maria. 1998. “Computerized Corpora and the Future of Translation Studies.” *Meta: Journal des Traducteurs/Meta: Translators’ Journal* 43(4): 652-659.

The Pervasiveness of Corpora in Translation Studies

Sara Laviosa[1] *Kanglong Liu*[2]

Address: [1]Department of Humanistic Research and Innovation, University of Bari 'Aldo Moro', Bari, Italy;
[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China
E-mail: [1]sara.laviosa@uniba.it; [2]kl.liu@polyu.edu.hk
Correspondence: Sara Laviosa
Citation: Laviosa, Sara, and Kanglong Liu. 2021. "The Pervasiveness of Corpora in Translation Studies." *Translation Quarterly* 101: 5-20.

Abstract

We take the view that, in order to appraise the advances made in any field of knowledge at a given point in time, it is important to revisit the past and view it through the lens of present achievements. Consequently, in this paper we trace the development of Corpus-based Translation Studies (CBTS) from its origin to the present day by surveying landmark publications and international events. We then make some recommendations for the future. Our aim is to show that the pervasiveness of corpus use in the pure and applied branches of Translation Studies is the result of a gradual process of integration of theory-driven and application-driven research. Indeed, the integration of translation theory, description and practice is, we believe, a hallmark of corpus research today and one of its main achievements. We contend that, if we are to make further advances in this area of scientific enquiry, we should endeavour to harmonize the concerns of professional translators, translator trainees, translator trainers, and translation scholars. The first author provides an overview of corpus use in Translation Studies from the early 1990s to the first decade of the new millennium. The second author assesses the state of the art of corpus studies of translation over the last ten years. In the concluding section, we jointly summarize the main points we covered throughout the paper and then look forward to the future.

1. Corpus-based Translation Studies: the beginnings

Let us begin our paper with some key definitions. Corpus-based Translation Studies (henceforth CBTS) is an area of research that adopts and develops the methodologies of Corpus Linguistics to analyse translation practices for theoretical, descriptive and applied purposes. Corpus Linguistics is an approach to the empirical study of language that relies on the use of corpora. Corpora are collections of authentic unabridged texts or whole sections of text held in electronic form and assembled according to specific design criteria "to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair 2005, 16).

Our overview of CBTS begins in 1993, when, in a paper published in a volume co-edited in honour of John Sinclair, Mona Baker expounded the rationale for applying the methods and tools of Corpus Linguistics to the study of translation (Baker 1993). More specifically, Baker outlined developments in Translation Studies that supported "a move towards corpus-based research" (Baker 1993, 236). The first was the replacement of the static notion of equivalence – traditionally viewed as formal correspondence of grammatical and syntactic structures – with the dynamic concept of functional equivalence between a source and a target text. The concept of functional equivalence shifted the focus of analysis from the source text vis-à-vis the target text (source-text orientation) to language text types and translated texts (target-text orientation).

Another development considered favourable to corpus-based research was the neo-Firthian view that meaning arises within a specific situational and linguistic context. This view links the concept of equivalence to usage rather than semantic meaning, and the study of usage requires the analysis of large quantities of authentic source texts and their translations. A third development that supported the use of corpora for the study of translation was the growing influence of polysystem theory in literary and translation studies. This theory, conceived by Itamar Even-Zohar in the late 1970s and developed by Gideon Toury in the 1980s and 1990s, views translated literature as a system in its own right, which interacts with other co-systems that constitute the whole target language literary polysystem.

Some fundamental changes were brought about by this novel theory. There was a shift away from the traditional analysis of individual source texts vis-à-vis their translations to the study of large numbers of translated texts. Translations were viewed first and foremost as "target language utterances" (Toury 1985, in Baker 1993, 239) capable, as such, of influencing the literary canons and language of their recipient culture. Moreover, polysystem theory asserts that translation is a creative, rather than a derivative, activity involving the adaptation of the source text to the target culture. Within this theory, norms are conceived as systematic choices made by translators at a particular time in a given culture. The concept of norms presupposes historical and cultural variation and is oriented towards the target, rather than the source culture. It also informs a concept of equivalence which is no longer prescriptive and absolute, but

descriptive and socio-culturally determined.

According to Baker, all these changes provided the ideal conditions for introducing and developing corpus-based research in the pure branch of Translation Studies in order to investigate: a) the universals of translation; b) the "operational norms" that constrain translational behaviour in a given socio-cultural context (Toury 1978, in Baker 1993, 246); c) the intermediate stages of the process of translation; d) the size and nature of the unit of translation; and e) the nature and limits of equivalence. With regard to the universals of translation, Baker (1993, 245) argued that a translated text is the "result of the confrontation of the source and target codes", and defined universals as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (243).

It is worth pointing out that the term 'universals' was not used by Baker with the intention of evoking associations with the concept of 'language universals' or 'universals of language', that was elaborated in linguistics by Noam Chomsky (1965) and Joseph H. Greenberg (1966). On the contrary, in Baker's paper and in Translation Studies generally the term 'universals' can be seen as an example of a rebranding concept, i.e. "the rebranding of the basic notion of a (widespread) tendency" (Chesterman 2019, 19). Since Baker's influential paper, these general tendencies have been explicitly defined in order to be falsifiable and empirically tested.

The first doctoral thesis that investigated translation universals by means of a corpus-based methodology was undertaken by the first author of the present paper under the supervision of Mona Baker and Juan C. Sager at the University of Manchester's Institute of Science and Technology (UMIST) (Laviosa-Braithwaite 1996). The study involved the design and compilation of the English Comparable Corpus (ECC) a monolingual multi-source-language comparable corpus of English literary and broadsheet newspaper texts. Firstly, the study formulated the interpretive hypothesis that lexical simplification can be viewed as the "process and/or result of making do with *less* [sic] words" (Blum-Kulka and Levenston 1983, 119, original emphasis). Then, a general testable descriptive hypothesis stated that, independently of source language and text type, translators working into English as language A tend to restrict the range of vocabulary available to them, and use a lower proportion of content words over grammatical words. Based on descriptive and inferential statistical analyses, the study showed three patterns of lexical use in both newspaper and literary texts. These patterns indicate that the range of lexical variety is narrower in translational English compared with non-translational English. In fact, translated texts display a lower lexical density, a higher proportion of high-frequency words over low-frequency words and a higher repetition rate of high-frequency words.

In this PhD thesis the compound term 'corpus-based translation studies' first appeared in the literature. It was later used by Miriam Shlesinger (1998) in an article entitled "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies", in which she launched the idea of applying a corpus-based methodology for descriptive interpreting studies.

The goal she envisaged for this body of research was to unearth the specificity of interpreting vis-à-vis original oral discourse and written translation in the same language.

After the completion of the first corpus study of lexical simplification (Laviosa-Braithwaite 1996), the quest for translation universals established itself as a line of enquiry pursued within corpus-based Descriptive Translation Studies (see Laviosa and Liu forthcoming 2021 for a review). The goal was to unveil the characteristics of the 'third code', as posited by William Frawley, who claimed that, since the act of translation involves bilateral consideration and accommodation of at least two codes, translation itself "emerges as a code in its own right, setting its own standards and structural presuppositions and entailments, though they are necessarily derivative" of the codes involved (Frawley 1984, 169).

An example of a study in search of the third code is Linn Øverås' (1998) investigation of explicitation in translational English and translational Norwegian using a bi-directional English-Norwegian parallel corpus of literary texts. The study aimed to unveil the specificity of the language of translation regardless of the contrastive differences between the two languages and attempted to draw some conclusions about the literary translational norms prevailing in the target culture. Øverås put forward a restricted descriptive hypothesis stating that English and Norwegian target texts tend to be more cohesive than their source texts. The results largely confirmed this hypothesis since the explicitating shifts, involving the addition and specification of lexical and grammatical items, were found to outnumber the implicitating shifts in both directions of translation, although English target texts displayed a lower level of explicitness vis-à-vis Norwegian target texts.

In the same special issue of *Meta* devoted to the corpus-based approach, where Shlesinger and Øverås published their corpus-based descriptive research into interpreting and translation respectively, other scholars presented their theoretical, empirical and application-driven scholarly work. Most notably, Baker (1998) expanded the agenda she had set out five years earlier. She discussed the need to develop a coherent corpus-based methodology for identifying the distinctive features of translational language. She contended that the aim of this research endeavour, which built upon the studies of scholars working within the descriptive and target-oriented perspective, would not merely be to unveil the nature of the 'third code' per se, but to understand the specific constraints, pressures and motivations that influence the act of translating intended as a mediated communicative event.

Furthermore, Sandra Halverson (1998) discussed the issue of representativeness in the design of general purpose translation corpora and provided a coherent theoretical framework within which data and methodology form a coherent whole to ensure the comparability of empirical findings. To this end, Halverson proposed a prototypical conceptualisation of the object category as opposed to a classical one. In this approach, the target population is regarded as a prototype category whose centre is occupied – but only for the cultures of industrialised western countries – by professional translations, whereas the peripheral positions are filled in by

clusters of different typologies, for instance, those carried out by trainees or non-professional translators or those between one's own best language and another language. The relationship between the centre and the periphery within the prototype is not one of inclusion or exclusion of the elements belonging to the category, but of resemblance. For the researcher this means that, in order to ensure representativeness, a sample corpus of the population of translated texts would have to be made up of an array of subcorpora having different degrees of significance and all being regarded as legitimate objects of study. Halverson acknowledged that prototypes are by definition culture-bound, so corpus-based findings cannot be generalised beyond the specific target population represented by a given corpus.

Adopting an interdisciplinary stance, Kirsten Malmkjær (1998) explained the advantages of using parallel corpora for contrastive and translation studies. For the contrastive linguists parallel corpora are valuable for investigating the differences and similarities in language use. For the translation scholar they are valuable for identifying translational norms. She then discussed two main problems connected with the use of parallel corpora for answering questions arising from within Translation Studies in particular. The first problem is that KWIC concordance lines do not always offer sufficient linguistic context to investigate features of whole texts. There exists, therefore, a risk that some aspects of translational behaviour may be revealed, while others may be overlooked. The second difficulty is related to the way parallel corpora are designed so as to include only one translation for each source text. This, as Halverson argued, may hide an important aspect of the translational phenomenon, namely the differences existing between the various translations of the same original work. To remedy these shortcomings, Malmkjær suggested complementing norm-oriented studies, which require large amounts of text, with smaller and carefully constructed corpora which consist of one source text and as many translations of it as possible, so that in-depth investigations of entire texts can be performed.

In addition to these reflections on methodological issues, other scholars presented the results of their empirical research. By way of example, Sara Laviosa (1998) reported on the findings of her doctoral study (Laviosa-Braithwaite 1996). Also, Jeremy Munday (1998) reported on the preliminary findings of the analysis of Edith Grossman's translation, *Seventeen Poisoned Englishmen*, of a short story by Gabriel García Márquez, *Diecisiete ingleses envenenados*. Munday used a variety of basic corpus linguistic analytical methods – word frequency lists, descriptive statistics and concordances – to explore texts inductively. Word frequency lists were first obtained for both source and target texts and then compared for identifying useful areas of investigation. Munday used intercalated text, i.e. a text obtained by manually keying in the translated text between the lines of the source text. He then ran concordances of this intercalated text and used them to carry out a contextualised comparative analysis of all the instances of selected lexical items in order to examine the shifts that build up cumulatively over the whole text as a result of the choices taken by the translator. Such analysis was carri-

ed out to understand the decision-making process underpinning the product of translation and infer the translator's textual-linguistic norms.

Munday's approach is therefore descriptive, product- and process-oriented and data-driven. He derived his hypotheses from observing differences that occur in the parallel frequency lists and during the manual construction of the intercalated text. These initial hypotheses were then investigated with the aid of additional automatic methods of analysis such as aligned concordance lines. Munday's investigation of the first 800 words of his full-text parallel corpus revealed shifts in cohesion and word order that occur over the whole text and have the effect of moving the narrative viewpoint from the first to the third person and thereby distancing the reader from the thoughts, experiences and feelings of the main character in the story.

With regard to applied corpus-based translation studies, the research carried out by Federico Zanettin (1998) and Lynne Bowker (1998) dealt with translator training. Zanettin demonstrated how small bilingual comparable corpora were useful to explore the stylistic features of a particular text genre by comparing words and phrases that have a strong formal resemblance (e.g. proper names and cognates) or are lexicographic translation equivalents. Zanettin provided concrete examples of such searches carried out with his students, who compiled an Italian-English comparable corpus of leading daily newspapers. The way in which President François Mitterand was talked about in the two languages, for example, presented interesting differences: *François Mitterand* or simply *Mitterand* was found to be commonly used in Italian, while English preferred *President Mitterand* or *President François Mitterand* or *Mr Mitterand*. Also, equivalent verbs typically used to introduce direct and reported speech were found to have different frequencies as well as syntactic and collocational profiles in the two languages. Even cognates such as *prezzi* and *prices* showed different collocational and colligational patterns. These data-driven learning investigations helped students to refine their knowledge of the source and target languages and develop their translation skills.

Still within a pedagogic perspective, Bowker addressed two main problems usually encountered by students training to become professional translators in specialized subject domains. One problem is the occurrence of terminological errors resulting from poor subject-specific knowledge. The other is the occurrence of errors due to a lack of specialized writing skills in the target language. Bowker's pilot study consisted in a translation experiment undertaken with a group of fourth-year undergraduate students at Dublin City University with English as their language A. They carried out two translations from French of two semi-specialized passages on optical scanners. One translation was completed with the use of bilingual and monolingual dictionaries together with non-lexicographic reference materials (e.g. manuals and brochures). The other translation was carried out with a bilingual dictionary and a 1.4 million-word specialized monolingual corpus of English articles on optical scanners, which was compiled from *Computer Select* on CD-ROM. The software used to analyse the corpus was *WordSmith* Tools.

The findings revealed that the corpus-aided translations were of higher quality in respect of subject field understanding (*sensibilité ayx* nuances was accurately rendered as *whatever their sensitivity to colour*); correct term choice (*vitre/glass paten or scan bed*); and idiomatic expression (*photodiodes sensible à la lumière/light-sensitive photodiodes or photosensitive diodes*). Bowker observed that, although there was no improvement with regard to grammar or register, the use of a specialized monolingual target corpus was not associated with poorer performance. The theoretical, descriptive and applied studies reviewed so far sowed the seeds of what would become a fully-fledged area of scholarly enquiry and practice in the new millennium, to which we now turn.

2. CBTS at the turn of the century

A series of important international initiatives marked the beginning of the newmillennium. The conference on "Research Models in Translation Studies", held at the University of Manchester's Institute of Science and Technology (UMIST) in 2000 and the 2001 European Society for Translation Studies (EST) congress in Copenhagen hosted panels devoted to corpus-based translation research. In 2001, Baker conducted a workshop in Pretoria to train South African researchers to build corpora for translation and interpreting research projects in various languages of South Africa. It was during that workshop that Alet Kruger, Kim Wallmach (University of South Africa) and Mona Baker (UMIST) had the idea of jointly hosting an international conference entirely devoted to CBTS in South Africa. The conference was held in Pretoria from 22 to 25 July 2003, and the title was "Corpus-based Translation Studies: Research and Applications". As Alet Kruger (2004, 2) wrote in the Editorial of the special issue of *Language Matters: Studies in the Languages of Africa*, where 19 selected papers presented at the conference were published,

> The aim of the conference was to consider ways in which corpora could be used to develop novel and challenging perspectives in the discipline, as well as ways in which they could support research outside the mainstream hegemonic research cultures.

The collection of papers was introduced by Laviosa (2004), who offered an examination of the relationship between CBTS and Descriptive Translation Studies, on the one hand, and CBTS and Corpus Linguistics, on the other. The aim was to establish which claims and predictions put forward in the early days of CBTS still held true and which were the most promising areas of CBTS research in the long term. The remaining papers were grouped into descriptive studies (literary and specialized texts), applied studies (translation and interpreting) and Bible translation. Monolingual comparable corpora and bilingual parallel corpora in different language combinations were used in this wide array of new studies in the field.

Also, in the early 2000s, two monographs were published in England (Laviosa 2002; Olohan 2004) together with the first collected volumes on corpus-based translator education (Zanettin et al. 2003) and translation universals (Mauranen and Pekka Kujamäki 2004). A second international conference devoted to CBTS was hosted in Shanghai from 31 March to 3 April 2007, "Conference and Workshop on Corpora and Translation Studies", and a new collected volume on the use of corpora in translator education was published in Europe (Beeby et al. 2009). Thirteen years on from the publication of the special issue of *Meta*, and seven years on from the publication of the special issue of *Language Matters*, another collection of papers on CBTS was published in England (Kruger et al. 2011). As the editors point out in the Introduction,

> The articles in this volume are written by many of the leading international figures in the field. They provide an overall view of developments in corpus-based translation (and interpreting) studies and also specific case studies of how the methodology is employed in specific scenarios, such as contrastive studies, terminology research and stylistics (Kruger et al. 2011, 1).

The lines of enquiry represented in this collection of papers were theory, description, applications, and tools. One of the novelties of this volume was the review paper by Robin Setton (2011) on Corpus-based Interpreting Studies (CIS). Setton demonstrated how the line of enquiry first proposed by Shlesinger (1998) was growing steadily thanks to the design of new software, access to larger, quality corpora of interpreted speeches, and new techniques for transcription, analysis, presentation and sharing of data. So, at the end of the first decade of the new millennium, CBTS emerged as an area of international research that was making inroads into the pure and applied branches of the discipline as a whole, it was offering new opportunities for the development of interpreting studies, and was adapting modern technologies to enhance theory, empirical research and practice for the benefit of translator and training and the work of the professional translator.

3. The state of the art of CBTS

As has been affirmed in the foregoing review, CBTS grew into a fully-fledged field of inquiry in less than two decades since Baker laid down the research agenda for CBTS. With the increasing sophistication of corpus tools and the emphasis of empirical research in Translation Studies, CBTS began to occupy a central role within the discipline as a whole and across adjacent disciplines. For example, the first interdisciplinary conference "Using Corpora in Contrastive and Translation Studies (UCCTS)" was launched by Richard Xiao in 2008 and has now become a biennial event which provides a forum for exploring the application of corpora in contrastive and translation studies.

The past decade, in particular, has witnessed the publication of an increasing number of books exploring various topics related to: corpus construction (Zanettin 2014); research methodology (Mikhailov and Cooper 2016); translator's style (Huang 2015; Mastropierro 2017); language contact through translation (Malamatidou 2017); and translation teaching (Liu 2020). We now outline some prominent developments of CBTS, most notably in the refinement and implementation of methodology.

Earlier debates centering on the notion and existence of translation universals has somewhat subsided and more and more researchers have cast their eyes on the methodological constraints of CBTS research. Unlike earlier studies, which were largely based on frequency-counting to decide whether an assumption could be supported or otherwise, recent studies have made it a norm to adopt statistical methods and techniques for hypothesis testing. Working from the field of CBTS, Hu (2016: 224-225) contends that

> the introduction of quantitative research in translation studies enables a researcher not only to conduct data-based statistical analysis of translated language, hence making translation research more scientific, but also to uncover translation regularities and translation norms unlikely to be generalized based on researcher's intuition and introspection.

CBTS researchers have been at the forefront in this 'quantitative turn' in translation studies, as evidenced by a number of representative monographs (e.g. Oakes and Ji 2012; Mikhailov and Cooper 2016) and a plethora of studies using advanced statistical methods. Instead of focusing on the sole variable of translation status, researchers have increasingly viewed translation as a language product shaped by a wide range of factors. In order to advance this rejuvenated research agenda, researchers have introduced "multifactorial design" (Kruger 2019) to the field of CBTS for uncovering the nature of translation as an activity directly or indirectly governed by linguistic, cognitive, socio-cultural or even political factors.

In corpus-based descriptive research, efforts continue to be made concerning the verification and falsification of translation universals with the increasing use of sophisticated statistical methods. The translation universals under investigation include: simplification (Liu and Afzaal 2021); normalization (Bernardini and Ferraresi 2011); and explicitation (Kruger and Van Rooy 2012). It is worth noting that in these studies the use of statistical methods for testing hypotheses and offering explanations has surpassed earlier research in terms of scientific validity and reliability.

Thanks to the methodological advances achieved in CBTS, researchers have also been able to explore the use of various linguistic indicators by looking across the disciplinary fence into the realm of computational linguistics. For example, the metrics of mean dependency distances (MDD) and dependency direction were used by Fan and Jiang (2019) to examine the simplification hypothesis. They found that translated English texts from Chinese are characterized by longer MDD and head-initial structures than original English texts.

Moreover, Hu and Kübler (2021) operationalized a number of entropy-based metrics informed by information theory to measure information density and complexity in translated texts to test the simplification hypothesis. Also, inspired by the work of Baroni and Bernardini (2006), more researchers have used data mining techniques to distinguish translation from non-translation (e.g. Lembersky et al. 2012; Ilisei 2013; Volansky et al. 2015). Overall, we have seen that this line of research has generated more solid data supporting the claim that translational language is categorically different from original writing.

Clearly, the research agenda on translation universals set out by Baker in the early 1990s has continued to have an impact on various disciplines and research fields. This demonstrates that translation studies continuously draws on concepts, methods and tools from across a range of neighbouring disciplines, integrating humanities and sciences into a comprehensive investigation of translation as a unique form of cross-cultural communication.

As the influence of CBTS continues to grow, more studies have been conducted to probe into many research areas which were barely touched upon in the previous two decades. In a recent collected volume, **Spring-cleaning: A Critical Reflection**, co-edited by María Calzada Pérez and Sara Laviosa, a number of researchers have specifically explored new research areas using CBTS methods, including: subtitling (Arias-Badia 2021); travel journalism (Brett, Loranc-Paszylk and Pinna 2021); and operatic audio description (Irene Hermosa-Ramírez). This shows that CBTS as a research area has the capacity to inform various fields of translation studies by fostering interdisciplinarity and empiricism. From the early quest of translation universals using comparable corpora, CBTS has now emerged as a truly interdisciplinary field of research encompassing a wide range of applications, approaches and objectives.

Corpus-based approaches have also been applied to the field of interpreting studies. While corpus-based interpreting studies (CIS) emerged at a later stage in comparison to corpus-based studies of written translation, it has continued to captivate the interest of researchers in recent years. Unlike written translation corpora, the compilation of interpreting corpora involves laborious work of speech-to-text transcription and manual annotation. This might be one of the reasons hampering the progress of this research area. Nonetheless, we have seen positive progress in using corpus methods to examine how interpreting language differs from other varieties of language outputs such as written translation and spontaneous speeches. Tang and Li (2016, 2017) examined the use of explicitation techniques by professional and trainee interpreters with the use of parallel corpora and found that the former tended to employ more explicitation than the latter, demonstrating that such techniques are closely related to the interpreter's competence.

Furthermore, based on the European Parliament Translation and Interpreting Corpus, Bernardini, Ferraresi and Miličević (2016) found that interpreting language features more simplification than translational language. Within the same line of enquiry, by comparing CI (Consecutive Interpreting) output with SI (Simultaneous Interpreting) output in a comparable corpus,

Lv and Liang (2019) found that CI output is more simplified than SI output. In all these recent studies, we can see that the earlier quest for translation universals has clearly made an impact on corpus-based interpreting studies (CIS), as evidenced by the vast number of studies examining the uniqueness of interpreting language. In the edited monograph titled *Making Way in Corpus-based Interpreting Studies* (Russo, Bendazzoli and Defrancq 2018), we can observe that CIS has emerged as a well-developed research area compared with a decade ago. CIS researchers have employed corpus technology to explore different aspects of interpreting, including the construction of the European Parliament (EP) interpreting and multimodal corpora (Bernardini et al. 2018), cognitive load in interpreting (Wang and Zou 2018) and interpreting universals or interpretese (Aston 2018). In sum, the development of CBTS has clearly provided an enormous impetus to the development of CIS. And now we are witnessing yet a new turn that is characterized by a fruitful exchange of theoretical insights, empirical data, as well as state-of-the-art methods and tools between CIS and CBTS. This new orientation is amply demonstrated by the papers contained in a collected volume that has just been published, as we are writing this article. It is titled **New Empirical Perspectives on Translation and Interpreting** and is co-edited by Lore Vandevoorde, Joke Daems and Bart Defranqc (2021), who, in their introductory chapter, declare their intention of "reuniting the sister disciplines of Translation and Interpreting Studies".

It is, therefore, not an overstatement to claim that enormous progress has been made in CBTS over the past decade. The use of corpora has increasingly been viewed as a flexible and useful methodology rather than a research precinct accessible to a limited number of so-called corpus researchers. Also, as a fully-fledged area of scholarly enquiry, CBTS has established itself as a mainstay across the three branches (theoretical, descriptive and applied) of Holmes' delineation of the field of translation studies.

4. Conclusion

This paper has traced the application and development of corpus methods in translation studies. Based on a review of important studies in the field, we can see that CBTS has attained great achievements not only in breadth (i.e. various research areas) and depth (i.e. fine-grained research methodology), but has also contributed to the development of translation studies as a whole. To a large extent, CBTS has significantly contributed to the establishment of Translation Studies as an independent discipline on merits of its scientific methods that have made it possible to investigate a variety of translational phenomena empirically. However, we should also be aware of the pitfalls of relying solely on data to confirm the obvious instead of "focusing on sense-making which follows the generation and presentation of statistical results" (Li 2017, 110). Similarly, we should also recall House's (2011, 206) statement that "[c]orpus evidence, and especially impressive statistics, should not be seen as an end in itself, but as a

starting point for continuing richly (re)contextualized qualitative work with values one finds interesting". In this respect, corpus use should be triangulated with other techniques. As foreseen by Laviosa a decade ago (2010, 86), the combination of corpus data, experimental, metatextual, ethnographic, and survey-based data will help contextualize, diversify and enrich linguistic evidence. Nonetheless, we are confident that corpus-based or corpus-assisted research grounded in data rather than unfalsifiable claims will continue to inspire translation scholars in their understanding and explanation of the translation phenomenon.

Finally, what does the future holds for CBTS in the light of the linguistic, socio-cultural, educational, technological, and professional changes that are taking place in postmodern societies? CBTS needs to offer a wide variety of corpus-based resources that represent not only major world languages but also languages of lesser diffusion and in different modalities (written, spoken, visual. auditory) in order to enrich our knowledge of the interrelationship between language and culture in different kinds of mediated communication. Also, CBTS needs to play a major role in promoting empirical and application-driven research that has a sound multidisciplinary theoretical foundation. Thirdly, CBTS needs to play a key role in bridging the gap between translation education (including translator training and pedagogic translation) and the professional world, so as to meet the needs of today's increasingly multilingual and globalized language industry.

References

Arias-Badia, Blanca. 2021. "Using Corpus Pattern Analysis for the Study of Audiovisual Translation: A Case Study to Illustrate Advantages and Limitations." *Monti* 13: 93-113.

Aston, Guy. 2018. "Acquiring the Language of Interpreters: A Corpus-based Approach". In *Making Way in Corpus-based Interpreting Studies*, ed. by Russo Mariachiara, Claudio Bendazzoli, and Bart Defrancq, 83-96. Singapore: Springer.

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies: Implications and Applications." In *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233-250. Amsterdam and Philadelphia: John Benjamins.

Baker, Mona. 1998. "Réexplorer la langue de la traduction: une approche par corpus." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43(4): 480-485.

Baroni, Marco, and Silvia Bernardini. 2006. "A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text." *Literary and Linguistic Computing* 21(3): 259-274.

Beeby, Allison, Patricia Rodrígues Inés, and Pilar Sánchez-Gijón. 2009. *Corpus Use and Translating*. Amsterdam and Philalelphia: John Benjamins.

Bernardini, Silvia, Adriano Ferraresi, and Maja Miličević. 2016. "From EPIC to EPTIC—Exploring Simplification in Interpreting and Translation from an Intermodal Perspective."

Target. International Journal of Translation Studies 28(1): 61-86.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard, and Bart Defrancq. 2018. "Building Interpreting and Intermodal Corpora : A How-to for a Formidable Task." In *Making Way in Corpus-based Interpreting Studies*, ed. by Russo Mariachiara, Claudio Bendazzoli, and Bart Defrancq, 21-42. Singapore: Springer.

Bernardini, Silvia, and Adriano Ferraresi. 2011. "Practice, Description and Theory Come Together - Normalization or Interference in Italian Technical Translation?." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 56(2): 226-246.

Blum-Kulka, Shoshana, and Eddie A. Levenston. 1983. "Universals of Lexical Simplification." In *Strategies in Inter-language Communication*, ed. by Claus Faerch, and Gabriele Casper, 119-139. London and New York: Longman.

Bowker, Lynne. 1998. "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43(4): 631-651.

Chesterman, Andrew. 2019. "Moving Conceptual Boundaries: So What?." In *Moving Boundaries in Translation Studies*, ed. by Helle V. Dam, Matilde Nisbeth Brøgger, and Karen Korning Zethsen, 12-25. London and New York: Routledge.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Fan, Lu and Yue Jiang. 2019. "Can Dependency Distance and Direction be Used to Differentiate Translational Language from Native Language?." *Lingua* 224:51-59.

Finbar Brett, David, Barbara Loranc-Paszylk and Antonio Pinna. 2021. "A Corpus-driven Analysis of Adjective/noun Collocations in Travel Journalism in English, Italian and Polish." *Monti* 13: 114-147.

Frawley, William. 1984. "Prolegomenon to a Theory of Translation." In *Translation: Literary, Linguistic and Philosophical Perspectives*, ed. by William Frawley, 159-175. Newark, University of Delaware Press.

Greenberg, Joseph H. ed. 1966. *Universals of Language*, 2nd edn. Cambridge, MA, and London: MIT Press.

Halverson, Sandra. 1998. "Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43(4): 494-514.

Hermosa-Ramírez, Irene. 2021. "The Hierarchisation of Operatic Signs through the Lens of Audio Description: A Corpus Study." *Monti* 13: 184-219.

House, Juliane. 2011. "Using Translation and Parallel Text Corpora to Investigate the Influence of Global English on Textual Norms in Other Languages." In *Corpus-Based Translation Studies: Research and Applications*, ed. by Alet Kruger, Kim Wallmach, and Jeremy Munday, 187-210. London: Continuum.

Hu, Hai, and Sandra Kübler. 2021. “Investigating Translated Chinese and Its Variants Using Machine Learning.” *Natural Language Engineering* 27: 1-34.

Hu, Kaibao. 2016. *Introducing Corpus-based Translation Studies*. Berlin: Springer.

Huang, Libo. 2015. *Style in Translation: A Corpus-Based Perspective*. Singapore: Springer.

Ilisei Iustina-Narcisa. 2013. *A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models*. PhD Thesis. University of Wolverhampton, UK.

Kruger, Alet, Kim Wallmach, and Jeremy Munday. 2011. “Introduction.” In *Corpus-Based Translation Studies: Research and Applications*, ed. by Alet Kruger, Kim Wallmach, and Jeremy Munday, 1-9. London: Bloomsbury.

Kruger, Alet, Kim Wallmach, and Jeremy Munday. eds. 2011. *Corpus-Based Translation Studies: Research and Applications*. London: Bloomsbury.

Kruger, Alet. 2004. “Editorial. Corpus-based Translation Research Comes to Africa.” *Language Matters. Studies in the Languages of Africa* 35(1): 1-4.

Kruger, Haidee, and Bertus Rooy. 2012. “Register and the Features of Translated Language.” *Across Languages and Cultures* 13(1): 33-65.

Kruger, Haidee. 2019. “That again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English”. *Across Languages and Cultures* 20(1): 1-33.

Laviosa Sara. 1998. “Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose.” *Meta: Journal des Traducteurs/Meta: Translators’ Journal* 43(4): 557-570.

Laviosa, Sara, and Kanglong Liu. forthcoming. “Translation Universals.” In *The Routledge Handbook on the History of Translation Studies*, ed. by Anne Lange, Daniele Monticelli and Chris Rundle. London and New York: Routledge.

Laviosa, Sara. 2002. *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam and New York: Rodopi/Leiden: Brill.

Laviosa, Sara. 2004. “Corpus-based Translation Studies: Where does it Come from? Where is it Going?.” *Language Matters. Studies in the Languages of Africa* 35(1): 6-27.

Laviosa, Sara. 2010. “Corpora.” *In Handbook of translation studies*, ed. by Yves Gambier, and Luc Van Doorslaer, 80–86. London: Routledge.

Laviosa-Braithwaite. 1996. *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. PhD Thesis. Centre for Translation and Intercultural Studies (CTIS). The University of Manchester, UK.

Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2012. “Adapting Translation Models to Translationese Improves SMT.” *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*: 255-265.

Linn Øverås. 1998. “In Search of the Third Code: An Investigation of Norms in Literary Translation”. *Meta: Journal des Traducteurs/Meta: Translators’ Journal* 43(4): 572-588.

Liu, Kanglong, and Muhammad Afzaal. 2021. “Syntactic complexity in translated and non-

translated texts: A corpus-based study of simplification." *Plos one* 16 (6): e0253454.

Liu, Kanglong. 2020. *Corpus-Assisted Translation Teaching*. Singapore: Springer.

Lv, Qianxi, and Junying Liang. 2019. "Is consecutive interpreting easier than simultaneous interpreting? – a corpus-based study of lexical simplification in interpretation." *Perspectives* 27 (1): 91-106.

Malamatidou, Sofia. 2017. *Corpus triangulation: Combining data and methods in corpus-based translation studies*. London: Routledge.

Malmkjær, Kirsten. 1998. "Love thy Neighbour: Will Parallel Corpora Endear Linguistics to Translators?." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43(4): 534-541.

Mastropierro, Lorenzo. 2017. *Corpus stylistics in Heart of Darkness and its Italian translations*. London: Bloomsbury Publishing.

Mauranen, Anna, and Pekka Kujamäki. eds. 2004. *Translation Universals: Do They Exist*. Amsterdam and Philadelphia: John Benjamins.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus linguistics for translation and contrastive studies: A guide for research*. London: Routledge.

Munday, Jeremy. 1998. "A Computer-Assisted Approach to the Analysis of Translation Shifts." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43(4): 542-556.

Oakes, Michael P., and Meng Ji. 2012. *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*. Amsterdam: John Benjamins Publishing.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London and New York: Routledge.

Russo, Mariachiara, Claudio Bendazzoli, and Bart Defrancq. 2018. *Making way in corpus-based interpreting studies*. Singapore: Springer.

Setton, Robin. 2011. "Corpus-based Interpreting Studies." In *Corpus-Based Translation Studies: Research and Applications*, ed. by Alet Kruger, Kim Wallmach, and Jeremy Munday, 33-75. London: Bloomsbury.

Shlesinger, Miriam. 1998. "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43(4): 486-493.

Sinclair, John M. 2005. "Corpus and text – Basic Principles." In *Developing Linguistic Corpora: A Guide to Good Practice*, ed. by Martin Wynne, 1-16. Oxford: Oxbow Books.

Tang, Fang, and Dechao Li. 2016. "Explicitation Patterns in English-Chinese Consecutive Interpreting: Differences between Professional and Trainee Interpreters." *Perspectives* 24(2): 235-255.

Tang, Fang, and Dechao Li. 2017. "A Corpus-based Investigation of Explicitation Patterns between Professional and Student Interpreters in Chinese-English Consecutive Interpreting." *The Interpreter and Translator Trainer* 11(4): 373-395.

Vandevoorde, Lore, Joke Daems, and Bart Defranqc. eds. 2021. *New Empirical Perspectives on Translation and Interpreting*. London and New York: Routledge.

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

Wang, Binhua, and Bing Zou. 2018. "Exploring Language Specificity as a Variable in Chinese-English Interpreting. A Corpus-based Investigation." In *Making Way in Corpus-based Interpreting Studies*, ed. by Russo Mariachiara, Claudio Bendazzoli, and Bart Defrancq, 65-82. Singapore: Springer.

Zanettin, Federico, Silvia Bernardini, and Dominic Stewart. eds. 2003. *Corpora in Translator Education*. Manchester: St. Jerome Publishing.

Zanettin, Federico. 1998. "Bilingual Comparable Corpora and the Training of Translators." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43(4): 616-630.

Zanettin, Federico. 2014. Translation-driven Corpora: *Corpus Resources for Descriptive and Applied Translation Studies*. London: Routledge.

Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-Translated French

Orphée De Clercq[1] *Gert De Sutter*[2] *Rudy Loock*[3]

Bert Cappelle[4] *Koen Plevoets*[5]

Address: [1,2,5] Department of Translation, Interpreting and Communication, Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium;
[3,4] Université de Lille, CNRS "Savoirs, Textes, Langage" CNRS Research Unit, 3 Rue du Barreau, 59650 Villeneuve-d'Ascq, France
E-mail: [1]orphee.declercq@ugent.be; [2]gert.desutter@ugent.be; [3]rudy.loock@univ-lille.fr; [4]bert.cappelle@univ-lille.fr; [5]koen.plevoets@ugent.be
Correspondence: Orphée De Clercq
Citation: De Clercq, Orphée, De Sutter, Gert, Loock, Rudy, Cappelle, Bert and Koen Plevoets. 2021. "Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French." *Translation Quarterly* 101: 21-45.

Abstract

This paper investigates the linguistic characteristics of English to French machine-translated texts in comparison with French original, untranslated texts in order to uncover what has been called "machine translationese". In the same vein as corpus-based translation studies which have focused on human-translated texts, and using a corpus-based statistical approach (Principal Component Analysis), we analyzed a ca. 1.8-million-word corpus of English to French translations of press texts, corresponding to the output of four machine translation systems: one statistical (SMT) and three neural (NMT) systems, namely DeepL, Google Translate, and the European Commission's eTranslation MT tool, in both its SMT and NMT versions. In particular, to complement a previous study on language-specific features in French (e.g. derived adverbs, existential constructions, coordinator et, preposition avec), a series of language-independent linguistic features were extracted for each text in our corpus, ranging from superficial text characteristics such as average word and sentence length to frequencies of closed-class lexical categories and measures of lexical diversity. Our results, which compa-

re the machine-translated data with a corpus of French untranslated data, allow us to uncover linguistic features in French machine-translated texts that clearly deviate from the observed norms in original French (e.g. average sentence length, n-gram features, lexical diversity), and which might serve as information for the post-editing process in order to optimize translation quality.

1. Introduction

Since the advent of neural machine translation (NMT) systems (Forcada 2017), it has become clear that the technology is disruptive and brings a lot of changes to the translation industry, both in terms of translation process and business model. In particular, as opposed to statistical machine translation (SMT), current NMT systems have started to provide output whose fluency can be quite impressive, although sometimes at the expense of accuracy or fidelity to the source text (e.g. Bojar et al. 2016; Macken et al. 2019). Such fluency has even led some researchers to claim that MT systems have reached "human parity" (Hassan et al. 2018), although such claims have been reassessed since (Toral et al. 2018). Nevertheless, the improved quality has led to the promotion of NMT in the field of translation (Daems and Macken 2019).

This makes it all the more crucial for translators to define their added value over the machine: they should develop their MT literacy, a concept defined by Bowker and Ciro (2019) for non-professionals, that is, they need to know what the machine can(not) do, what the difference is between human translations and MT output, and what to focus on during the post-editing (PE) process.

With NMT, the evaluation of MT systems has become a central issue, not only for the industry but also for translation training. The focus on fluency makes errors more difficult to identify (e.g. Castilho et al. 2017a, 2017b; Yamada 2019), and translators need to be provided with useful information for the PE process. A lot of debates have been taking place on the best way to assess MT output quality: use of metrics, human evaluation, or a linguistic evaluation with a corpus-based approach. This paper focuses on such a linguistic evaluation of MT output, through the analysis of English to French machine-translated texts produced by four different MT systems, in comparison with original, untranslated French data. Our analysis aims to explore different aspects of what has been called "machine translationese" (see e.g. Daems et al. 2017) by comparing machine-translated with original texts, relying on a corpus-based approach typical of corpus-based translation studies, with analyses carried out on a series of texts rather than a series of isolated sentences.

To this purpose a corpus of press texts was collected comprising both original (French) French and (British) English text material. The English data was translated into French using four different MT systems: DeepL, Google Translate – both NMT – and the European Com-

mission's eTranslation tool, in both the SMT and NMT flavor. This allowed us to compare the frequencies of a series of linguistic features in original vs. machine-translated French, with the same methodology and statistical techniques that have proven capable of distinguishing between student and professional translations (De Sutter et al. 2017). The final aim is to define the 'gap' that exists between machine-translated texts and the norms expected in untranslated texts, providing information on how to improve translators' invisibility as expected by today's market, thanks to a list of elements to focus on during the PE process.

Specifically, this paper aims to complement a previous study (Loock 2018, 2020), which analyzed the same data (with the exception of Google Translate output) by focusing on language-*specific* linguistic features (see section 2.2 for a summary and list of the features). Here our focus is on language- *independent* features like average word or sentence length, frequencies of part-of-speech (POS) tags, or frequencies of n-grams. These features are exploratively analyzed with Principal Component Analysis and the differences between original and machine-translated French are then tested by means of ANOVA. The analysis is therefore more sophisticated than in Loock (2018, 2020) and also includes the most famous publicly available MT system.

The remainder of this paper is structured as follows. In section 2 we describe related work on machine translationese within research on the quality of MT output and provide a summary of Loock (2018, 2020), of which the current study is an extension. Section 3 describes our methodology: corpus material, feature extraction, and statistical technique. Section 4 is dedicated to the presentation and discussion of the results, first for a general comparison between French machine-translated and untranslated texts, then for a finer-grained comparison of relevant linguistic features, and finally for a possible link with interference from the English source texts.

2. Related work

2.1 MT output evaluation and machine translationese

A lot of research has been devoted to the evaluation of MT output (see Moorkens et al. 2018 for a recent overview), in particular since the advent of NMT. Alongside metrics like BLEU (BiLingual Evaluation Understudy; Papineni et al. 2002) for example, researchers have relied on human evaluations to try and tackle the limits of automatic evaluation (see e.g. Babych 2014)[1]. For example, MT output has been evaluated by identifying and classifying errors (e.g. Federico et al. 2013; Van Brussel et al. 2018), measuring the amount of post-editing effort (e.g. Bentivogli et al. 2016), or ranking machine-translated texts by (non-)professionals (e.g. Bojar et al. 2015). Researchers have also focused on the identification of linguistic differences between machine-translated texts and original, untranslated texts in the same language.

Following the path mapped out by corpus-based translation studies (CBTS) with Baker (1993) as a starting point, which allowed for the identification of linguistic differences between original and (human) translated texts (see Laviosa 2002; Olohan 2004 or De Sutter et al. 2017 for a series of quantitative studies), some studies have identified linguistic differences between untranslated language and machine-translated language, post-edited or not, for series of sentences (e.g. Isabelle et al. 2017) or full texts (e.g. Vanmassenhove et al. 2019). In the latter case, machine-translated texts are gathered as electronic corpora, to be investigated with the quantitative methods of corpus linguistics. Just as some of the analyses of human-translated language in CBTS led to the identification of translationese (Gellerstam 1986) [2], the observation of machine-translated texts has led to the identification of so-called "machine translationese" for raw MT output and "post-editese" for post-edited MT output (MTPE). For example, Vanmassenhove et al. (2019) have shown that MT texts show lesser lexical variety than both original and human-translated texts for English to French and English to Spanish translations; Lapshinova-Koltunski (2015) has investigated English to German translations and has, in the same vein as what has been done for the analysis of human-translated texts, investigated the possible influence of so-called translation universals like simplification, explicitation, and normalization, by measuring lexical density/variety, the frequency of cohesion markers or specific grammatical categories (nouns vs. verbs). Similarly, Daems et al. (2017) and Toral (2019) have analyzed MTPE texts and shown the existence of post-editese, qualified by Daems et al. (2017) as "exacerbated translationese". In the present study, our focus is on raw, non post-edited MT output, and our aim is to check for the existence of machine translationese.

2.2 Language-specific vs. language-independent linguistic features

In order to uncover machine translationese or post-editese, researchers can focus on language-specific or language-independent linguistic features. For example, lexical variety or density in Lapshinova-Koltunski (2015) and average sentence/word length in Daems et al. (2017) are language-independent features, while Isabelle et al. (2017) focus on language-specific features for the evaluation of French to English MT output (e.g. verb-tense concordance, insertion of words like *fact or how*).

Our study investigates language-independent features (see complete list in section 3.2) in EN-FR machine-translated texts, as it is meant to complement a previous study on English to French machine-translated texts (Loock 2018, 2020) which used the same data (with the exception of the Google Translate subcorpus) and investigated specific linguistic features in French: the use of the hypernyms *chose and dire* ('thing' and 'say'), the coordinator *et* ('and'), the preposition *avec* ('with'), derived adverbs ending in *-ment* (the equivalent of *-ly* adverbs), and *il y* a existential constructions (the equivalent of *there is/are* constructions). The analysis of the EN-FR machine-translated texts (obtained by means of DeepL and eTranslation in both its SMT and NMT versions) has shown that, on an almost systematic basis, these specific lin-

guistic features show highly significant differences between original, untranslated French and machine-translated French from English, with much higher frequencies in machine-translated texts. These French linguistic features were selected as they are considered to be translational equivalents of the corresponding English items but these items' use in original English and original French shows differences in terms of frequencies. As the items are more frequent in original English than original French, one would expect the differences observed in machine-translated French texts to be the result of direct transfers between the English source texts and the French translated texts. However, the qualitative analysis carried out in Loock (2020) shows that this is not the case. Source language interference cannot fully explain the data, as we also notice differences in frequencies between the English source texts and the French machine-translated texts: for example, the frequency of *il y* a existential constructions in machine-translated French, higher than in untranslated French, is lower than that of *there is/are* constructions in the English source texts, suggesting that only some of them are translated literally (this is confirmed by the qualitative analysis of a sample of the corpus in Loock 2020).

3. Method

Using a corpus-based statistical approach, our objective is to investigate the existence of "machine translationese" for the language pair English-French. This approach consists of three steps. First, original French and English press texts are collected, after which the English texts are machine-translated using four well-known MT engines (section 3.1). Next, all corpora are preprocessed (including tokenization, lemmatization and part-of-speech tagging) and subsequently a series of linguistic language-independent features are extracted (section 3.2). In the third step, multivariate statistical analysis techniques are performed and the output is analyzed (section 3.3).

3.1 Data collection

The data used for this study contains three kinds of texts: (i) original texts written in (British) English, (ii) their translations into French by means of 4 different MT systems, and (iii) untranslated texts written in (French) French. Both series of original texts (i/iii) are extracted from the TSM press corpus (*Traduction Spécialisée Multilingue* corpus), an open-ended corpus compiled at the University of Lille for a comparative grammar class in a master's programme (Loock 2019). The corpus is a comparable corpus containing original, untranslated press texts taken from quality newspapers in British English (e.g. *The Guardian, The Observer, The Times*), American English (e.g. *The New York Times, The Wall Street Journal*), and (French) French (e.g. *Le Monde, Libération, La Voix du Nord*), with different news domains being covered: business and finance, crime, culture, environment, health, international news,

politics, science & technologies, sports and travel. At the time the current study was initiated, the TSM corpus contained ca. 1.6 million words (2.4 million words today). Table 1 provides a description of the version of the TSM corpus used for the present study. All French texts were used (1,440 texts amounting to 833,590 words); for English the British English subcorpus was selected (490 texts amounting to 374,326 words).

Table 1: Content of the TSM press corpus

	Ori US_EN	Ori GB_EN	Ori FR
Business & Finance	27 487	6 136	64 361
Crime	44 315	43 710	120 343
Culture	30 570	46 839	107 080
Environment	41 500	32 367	101 924
Health	34 790	28 170	78 022
International News	33 767	29 168	91 147
Politics	45 840	46 901	127 500
Science & Technologies	45 269	47 213	94 391
Sports	45 156	43 766	125 033
Travel	40 748	50 056	108 493
Total #tokens	**389 442**	**374 326**	**833 590**
#texts	**437**	**490**	**1 440**
GRAND TOTAL	**1 597 358**		

As far as machine-translated texts are concerned, the 490 British English texts were translated using four translation engines: DeepL, Google Translate – two commercial neural engines that are freely available online – and the engine developed by the European Commission's Directorate-General for Translation, called eTranslation, both in its SMT and NMT versions. DeepL [3] is trained on the corpus used for the Linguee website[4] and has become known for the quality of the target language, sometimes at the expense of accuracy (Bojar et al. 2016). Google Translate [5] is probably the most well-known generic MT tool and is frequently the object of scientific studies on the quality of MT output. Both tools have been providing internet users with neural machine translations since NMT went mainstream (around 2016). The eTranslation tool[6], both in its SMT and NMT flavors, has been designed for internal use at the European Commission and is not available to the general public,[7] although public administrations as well as small and medium-sized enterprises can currently make use of it, with September 2018 marking the arrival of the NMT version for the EN-FR language pair. In spite of its confidential nature, eTranslation has been the object of a few studies (Macken et al. 2020; Rossi and Chevrot 2019; Loock 2020).

The translations were retrieved in spring 2018 for DeepL and eTranslation SMT, Decem-

ber 2018 for eTranslation NMT, and July-August 2019 for Google Translate. Each of the 490 texts was translated individually, by copying/pasting the text online or by uploading the different files. Table 2 provides some corpus statistics of the machine-translated data, the British English source texts as well as a specification of the size of the comparable original French corpus used for the present study.

Table 2: Content of the corpus used for this study

Subcorpus		#texts	#tokens	Abbreviation
Original French		1 440	833 590	ORI_FRA
Machine-translated French	DeepL	490	442 439	DeepL
	eTranslation NMT	490	451 704	eNMT
	eTranslation SMT	490	445 914	eSMT
	Google Translate	490	431 297	GoogT
Original English (British)		490	374 326	SCR_ENG
GRAND TOTAL		3 890	2 979 270	

It should be noted that the original texts from the TSM press corpus belong to the press genre, while none of the 4 MT tools used have been trained or optimized for the translation of such texts (DeepL and Google Translate are generic MT tools; eTranslation has been trained on institutional data). This is of course a limitation of our study, since none of the 4 MT systems have been trained to translate press texts specifically, meaning the tools are not fully fit-for-purpose.

3.2 Data processing

A number of language-independent features have been extracted from the different subcorpora. For this extraction, it was crucial to linguistically preprocess all three corpora. This preprocessing consisted of three steps: tokenization, lemmatization and part-of-speech (POS) tagging. The LeTs preprocessing toolkit (Van de Kauter et al. 2013) was used for this purpose. The complete list of 22 features is presented in Table 3.

As can be derived from this table, the list contains two basic readability features (average word and sentence length), measures of lexical creativity and originality (e.g. type-token ratio, lexical density, hapax legomena), basic frequency information on different part of speech categories (lexico-grammatical features) and features indicating the degree of syntagmatic patterning or formulaicity (3- and 4-grams). All features are extracted at the text level; the legomena and ngram features are calculated by comparing an individual text with a background corpus. For example, for the legomena features we count how many French words in a certain text also occur one (hapax), two (dis) or three (tris) times in the entire French corpus used for this study. For the ngram features we check the number of combinations of three (3-gram) or

Table 3: All 22 language-independent features which were extracted from every text.

Feature type	Feature name	Abbreviation
Readability	Average sentence length	ASL
	Average word length	AWL
Lexical creativity	Lexical density	Den
	Type-token ratio	TTR
	Hapax legomena	Hapax
	Dis legomena	Dis
	Tris legomena	Tris
Lexico-grammatical	Frequency of common nouns	NOM
	Frequency of proper nouns	NAM
	Frequency of adjectives	ADJ
	Frequency of adverbs	ADV
	Frequency of verbs	VER
	Frequency of pronominals	PRO
	Frequency of determiners	DET
	Frequency of foreign words	FW
	Frequency of interjections	INT
	Frequency of numerals	NUM
	Frequency of prepositions & conjunctions	KON_PRP
Formulaicity	3-grams (word)	N3_wrd
	3-grams (POS)	N3_pos
	4-grams (word)	N4_wrd
	4-grams (POS)	N4_pos

four (4-grams) words or part-of-speech categories belonging to the 100 most frequent combinations in the corpus. Please note that when we refer to words for the legomena and ngram features, we actually mean the lemmas. Regarding the lexico-grammatical features it should be noted that the morphosyntactic categories prepositions and conjunctions were merged into one feature as we could not unequivocally distinguish these part-of-speech categories in the tagsets of both languages.

3.3 Statistical analyses

All language-independent features were extracted by means of custom-made Python scripts. This resulted in a data matrix in which every row contains the numerical information of the 22 features with respect to a given text. Every text is thus represented as a feature vector consisting of the scores of 22 linguistic features. Moreover, the origin of the text is also taken into

account – original French, machine-translated French using either DeepL, GoogleT, eNMT or eSMT or original English – leading to a 23-dimensional vector.

After having extracted this quantitative information from the corpora, we used Principal Component Analysis (PCA) to inspect the correlation structure of our data matrix in a lower-dimensional structure (see the seminal work by Baayen (2008) for more information, and Jenset and McGillivray (2012) or Evert and Neumann (2017) for examples of use of this methodology to uncover differences between original and translated language). For ease of presentation, we will present only two-dimensional plots in the remainder of this paper. These visualizations will reveal whether original and machine-translated French differ from each other, which could hint at machine translationese, and if so, whether all MT engines present the same picture. By also incorporating the corpus of British English source texts, possible shining-through from the source texts might also become apparent. This first explorative analysis is subsequently corroborated by an ANOVA of each linguistic feature, where the difference between original French, machine-translated French (with each of the four MT engines) and English is statistically tested. We will only report on the ANOVAs of the features which yield explicit differences.

4. Results and discussion

As mentioned in section 3, all texts from each subcorpus were first preprocessed, after which 22 language-independent features were extracted. Next, PCA was used to analyze the data. The results of this PCA are presented in Figure 1.

4.1 PCA results

Before interpreting this plot, one should note that the abbreviations printed in black represent the 22 linguistic features (see Table 3) and that each colored item represents a text coming from one of the different subcorpora. The numerical values on the x- and y-axis do not have a straightforward interpretation; what is meaningful, however, is the relative position of the different items vis-à-vis each other and vis-à-vis the linguistic features in the plot: the closer two items are, the more similar their linguistic characteristics (and vice versa); when a text is close to a given linguistic feature this means that the feature is clearly present in this text.

Given the number of texts, subcorpora, and features, this biplot is rather difficult to read. However, what immediately draws the attention is that we can distinguish between the English source texts (yellow) at the top, the original French texts (red) in the middle and the machine-translated French texts (all the other colors) more at the bottom. This means that there exist some clear differences between the different corpora, and that the linguistic characteristics of French machine-translated texts show some differences with both original French texts and

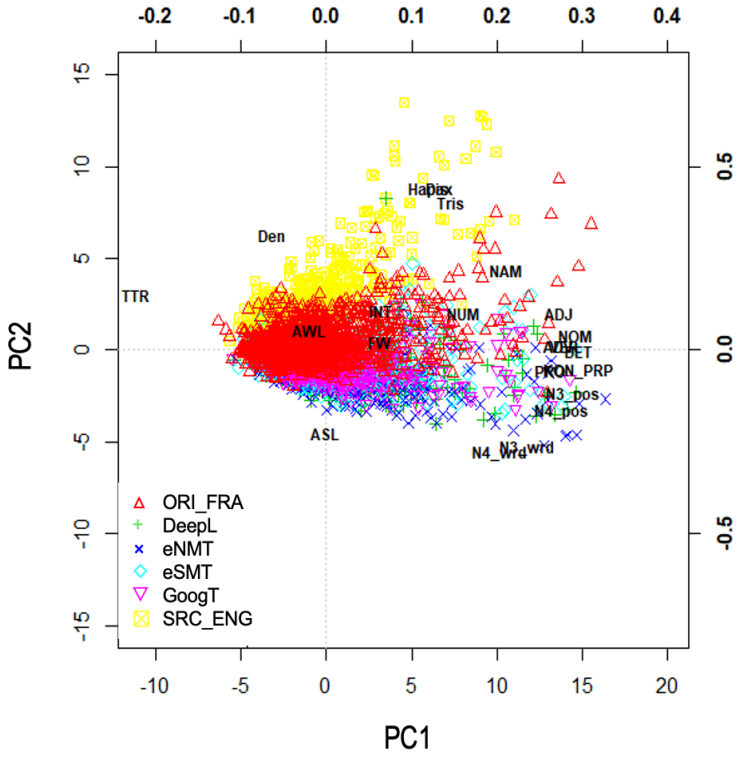


Figure 1: Biplot of the PCA for the original French (red triangle), the machine-translated French (green plus sign: DeepL, darkblue cross: eNMT, light blue diamond: eSMT, pink triangle: GoogT) and the British-English source texts (yellow check-marked square)

English source texts.

In order to better focus on the variation between the different varieties of French (untranslated texts and machine-translated texts for the different MT tools), we provide Figure 2, which depicts the same PCA analysis but with the visualization of the English source texts left out.

Looking at the original versus machine-translated French texts, there is quite some overlap, represented by the big colored blobs in the middle, though overall we also observe that certain features seem to pull down the machine-translated texts towards the right bottom, indicating that there does exist such a thing as machine translationese. If we look closer at which features cause this we see that it is mostly due to the average sentence length (ASL) and ngram features (N3_wrd, N4_wrd, N3_pos, N4_pos).

4.2 ANOVA analyses

This is where ANOVA analyses can shed more insights, as these test for each feature individually whether there is a significant difference between the different settings.

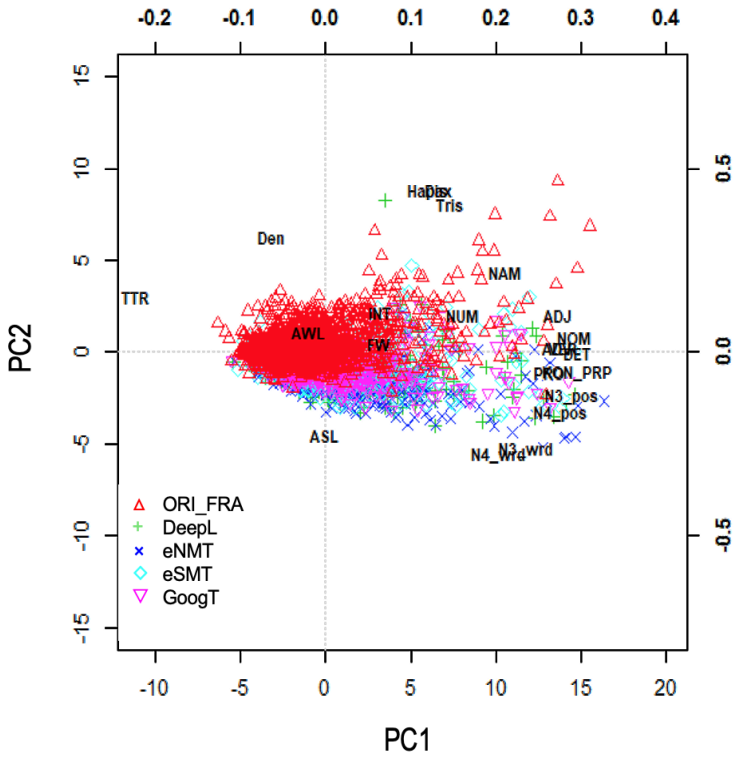


Figure 2: Same biplot as Figure 1 but without visualizing the British-English source texts.

4.2.1 Average sentence length

Figure 3 presents the ANOVA analysis for average sentence length. In this and all subsequent ANOVA graphs, interval plots are shown for every feature versus all settings, with dots representing the mean and pink lines the confidence intervals. If the intervals in different settings are far away from each other this indicates a substantial difference between the settings and if there is no overlap at all, this difference is statistically significant.

What this ANOVA reveals is that MT creates longer sentences in comparison with what is expected in French (ORI_FRA). We observe a sentence length increase of 18.2% (DeepL), 20.6% (eNMT), 19.1% (eSMT) and 15.2% (GoogleT). This is similar to human English to French translation, where a sentence length increase of around $20 - 25\%$ is considered to be normal[8].

4.2.2 ngram features

The ANOVAs of the different ngram features are depicted in Figures 4 a/b/c/d. These features are based on the top-100 most frequent combinations of 3 or 4 words (lemmas) or part-of-speech categories in both languages. Table 4 presents the top five combinations of each ngram feature in the entire French corpus.

Regarding the word ngrams both trigrams and fourgrams indicate a pronounced diffe-

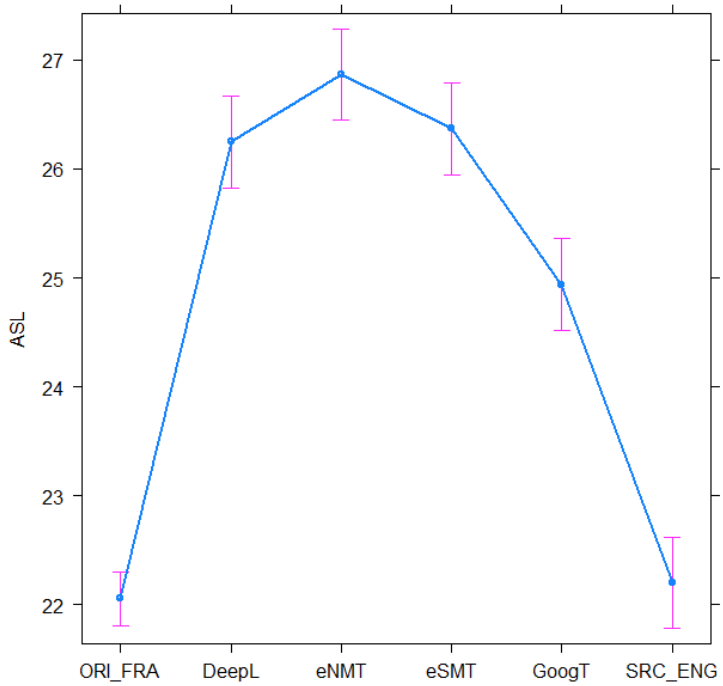


Figure 3: ANOVA average sentence length (ASL)

Table 4: Top five combinations of the French ngram features used for this study, combinations of 3 (N3_wrd) or 4 (N4_wrd) lemmatized word forms and of 3 (N3_pos) or 4 (N4_pos) part-of-speech combinations.

N3_wrd	N3_pos	N4_wrd	N4_pos
ne être pas	PRP DET NOM	*il se agir de*	NOM PRP DET NOM
il y avoir	DET NOM PRP	*Il ne y avoir*	PRP DET NOM PRP
ne avoir pas	NOM PRP NOM	*ce ne être pas*	DET NOM PRP NOM
le un du	NOM PRP DET	*se agir de un*	DET NOM PRP DET
ce être un	VER DET NOM	*avoir déclarer que le*	VER PRP DET NOM

rence between original and machine-translated French. All translation engines rely more on common words combinations and standard phrase structures. Let us have a closer look at the top three most and least frequent word ngrams in original French, compared to their position in the machine-translated texts, as presented in Table 5. The numbers represent the index of this ngram in the list of all 100 ngrams for each setting. The closer the colour is to red, the less frequent, the closer the colour is to blue, the more frequent[9]. From this table it can clearly be deduced that regarding the top three most frequent ngrams the MT engines are more or less

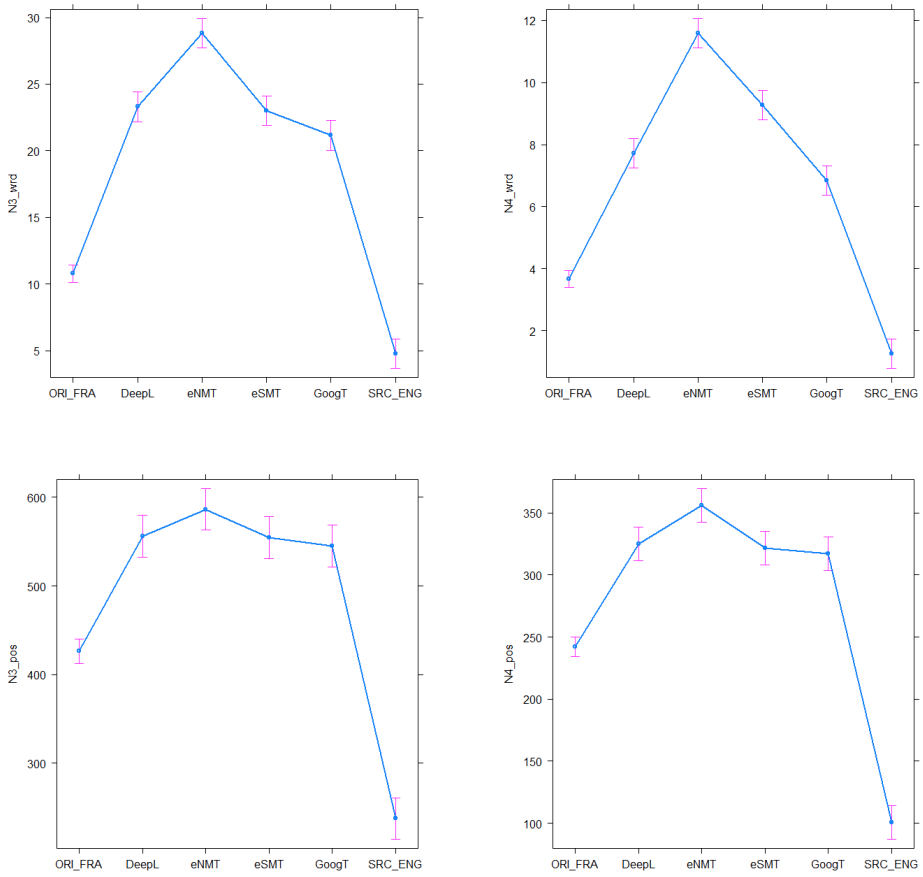


Figure 4: a/b/c/d. ANOVA analyses of the ngram features, based on combinations of 3 or 4 words (N3_wrd, N4_wrd) or part-of-speech tags (N3_pos, N4_pos).

in line with original French, especially the trigrams. However, when considering the top three least frequent ngrams, we observe that the MT engines use these ngrams much more frequently than in original French. Again this finding is more pronounced for the trigrams than for the fourgrams.

There are also quite some individual differences among the engines: for the trigrams especially eNMT uses more frequent combinations and for the fourgram features all MT engines differ significantly from each other, with the eTranslation systems standing out most clearly from the others. Inspecting the POS ngrams, there is no pronounced difference among the different MT engines for both the trigrams and fourgrams. However, machine-translated French also here clearly relies more on the same combinations of POS-tags than original French. We could say that the machine translation engines tend to "play safe". This is in line with the normalization translation universal, already found in human translations and defined by Baker

Table 5: Top three combinations of the most and least frequent lemmatized word form ngrams in the original versus the machine-translated French. The color range represents the frequency: the closer to red, the less frequent an ngram is; the closer to blue, the more frequent.

	ORI_FRA	DeepL	eNMT	eSMT	GoogT
Trigrams					
ne être pas	1	1	1	1	1
il y avoir	2	2	2	3	3
ne avoir pas	3	4	3	2	4
déclarer que il	100	74	31	32	26
déclarer que le	99	14	14	89	8
avoir déclarer que	98	3	6	13	2
Fourgrams					
il ne y avoir	1	8	3	4	6
ce ne être pas	2	5	21	12	2
il se agir de	3	2	1	3	4
avoir déclarer que il	97	19	19	15	3
du pays de Galles	96	62	72	63	36
avoir déclarer que le	95	1	4	38	1

(1993) as the exaggeration of features in the target language and conformity to its typical patterns. Given that the MT engines are being trained on large amounts of parallel human-translated data, then maybe this is why a normalization effect can also be found in MT texts. Moreover, MT engines relying more on the same word combinations also corroborates the work by Van Massenhove et al. (2019) which found that the inherent nature of data-driven MT systems to generalize over the training data has a quantitatively distinguishable negative impact on word choice, leading to less lexical diversity and bias. This is referred to as "algorithmic bias", characterized by an "exacerbation of dominant forms" (Van Massenhove et al. 2019, 223).

4.2.3 Lexical diversity

Given the findings in the previous subsection we would expect that when looking at measures of lexical diversity the MT engines reveal a similar tendency, namely of being lexically less diverse. The ANOVA plots presented in Figure 5 indeed confirm this hypothesis.

Whereas the interval plots of the original French and source English texts are similar to each other, they are more elevated than the type-token ratios present in the machine-translated texts, suggesting that machine translations are lexically less diverse. This corroborates previous similar findings by, for example, Toral (2019) or Vanmassenhove et al. (2019), which

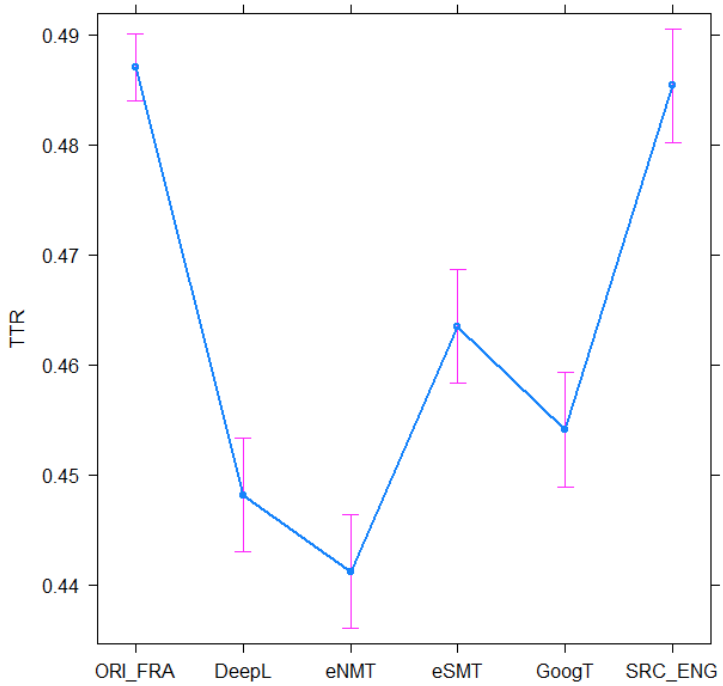


Figure 5: ANOVA analysis representing the Type-token ratio (TTR)

revealed that MT output has lower lexical diversity than human translation and that both have a lower lexical diversity than human-written, naturally composed text in the same language.

In this respect, it is also interesting to consider the legomena features, which are presented in the ANOVAs in Figures 6 a/b/c.

Overall, original French exhibits more hapax, dis and tris legomena than machine-translated French, which also hints at a difference in lexical diversity. The English source texts even comprise a much higher number of all three types of legomena than original French and one could thus expect some influence of this in the MT texts. However, this is not the case, which further underlines MT's incapability to produce lexically diverse translations independently of the lexical diversity in the source texts.

When comparing the four different MT engines we observe more or less the same tendencies for dis and tris legomena; however, the hapax legomena exhibit more pronounced differences, especially between the eTranslation systems on the one hand and DeepL and GoogleTranslate on the other hand. The eTranslation systems produce many more hapax legomena; more specifically, each text translated with the European Commission's translation engines has on average 6.25 (NMT) and 7.92 (SMT) hapax legomena per text versus 3.43 and 2.59 for DeepL and GoogleTranslate, respectively. By comparison, the average number of hapax

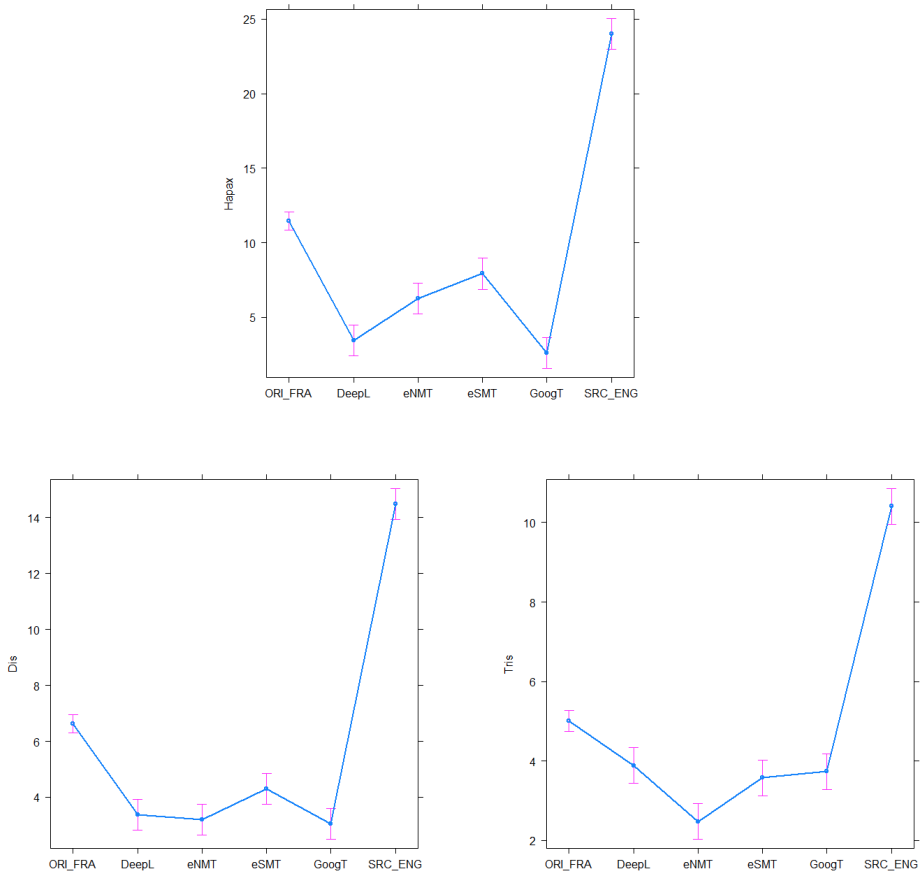


Figure 6: a/b/c. ANOVA analyses of the legomena features representing Hapax, Dis and Tris legomena

legomena in original French texts amounts to 11.52. This could imply that the output of the eTranslation systems are closer to original French when it comes to hapax legomena.

To get more insight into this, we inspected the average number of hapax legomena within each news domain (Table 6). As expected, we observe overall higher numbers for the eTranslation engines, and especially the SMT engine, for which numbers are closest to original French. However, what also draws the attention are the higher numbers – indicated in bold – in two different news domains, namely *Culture* and *Travel*, in all French corpora. This probably hints at more creative and unique language use in these two domains.

In order to get more insights into which hapax legomena are produced by the different MT engines in such domains, we manually inspected two texts, one of the *Culture* and one of the *Travel* domains, with an elevated number of hapax legomena [10]. Table 7 presents the results of this analysis, each cell containing a percentage of hapax legomena corresponding to

Table 6: Average number of hapax legomena per text in every news domain in the original French texts (ORI_FRA) compared to the machine-translated text with the four different MT engines

News domains	ORI_FRA	DeepL	eNMT	eSMT	GoogT
Business	10.09	2.08	2.5	5.25	1.67
Crime	6.49	2.05	4.05	4.79	1.69
Culture	**18.42**	**5.35**	**11.84**	**14.04**	**4.31**
Environment	11.07	2.85	3.51	5.76	2.68
Health	9.94	1.85	3.64	4.92	1.69
International News	8.5	2.25	3.52	4.96	1.85
Politics	8.75	2.44	2.24	3.7	1.54
Science & Technologies	11.19	2.79	5.75	7.91	2.61
Sport	11.63	5.87	7.24	8.56	2.24
Travel	**25.94**	**6.52**	**18.2**	**20.73**	**5.93**

one of the following categories:

- **Existing:** refers to French existing words, i.e. listed in a dictionary. Examples are *diabolisé, archétypal* or *microfissuré*.

- **Understandable:** refers to words which are not "official", but which are easy to process or understand. Examples are *zombifié, cavalage* or *vampiriquement*.

- **English:** refers to words that were not translated but merely copied from English. Examples are *avid, zombified* or *torch-lit*.

- **Made-up:** refers to made-up words which are hard to understand or words which were slightly adapted from English to French standards. Examples are *vortir, tonnamment, bienbouffeur* or *torch-éclairé*.

Table 7: Percentage of hapax legomena belonging to one of the four categories, calculated separately for every MT engine

	DeepL		eNMT		eSMT		GoogleT	
	Culture	Travel	Culture	Travel	Culture	Travel	Culture	Travel
Existing	**75.0**	**80.0**	31	28.0	24.0	32.5	**77.0**	**62.5**
Understandable	12.5	0.0	0.0	0.0	0.0	0.0	15.0	12.5
English	0.0	10.0	11.0	8.0	**72.0**	**52.5**	0.0	12.5
Made-up	12.5	10.0	**57.0**	**64.0**	3.0	15.0	8.0	12.5

The highest percentages per category are indicated in bold, and we clearly observe that both DeepL and GoogleT mostly produce unique words (hapax legomena) which exist in French. The same cannot be said for the eTranslation engines: the eNMT engine produces ma-

ny made-up words whereas the eSMT engine leaves many English words untranslated. This analysis contradicts what the numbers of Table 6 suggested: eTranslation tools and especially the eSMT engine are *not* closer to original French when it comes to hapax legomena.

Especially the made-up words are a problem to be mitigated, as research on reading comprehension of NMT nonsense words has found that this deteriorates comprehension and also leads to less confidence among readers; on the contrary, comprehension questions on words that are left untranslated are often answered more correctly (Macken et al. 2019). If we consider Table 7, especially the eNMT engine produced many nonsense words.

4.2.4 Verb, common noun, and proper name frequency

We conclude our discussion of the results by presenting one ANOVA where the French machine-translated texts seem to exhibit interference from the English source texts. Figure 7 presents the ANOVA of the frequency of the part-of-speech category verbs. Here, we observe a clear difference between original French and each of the machine-translated French corpora, which, in turn, are closer to the source English corpus.

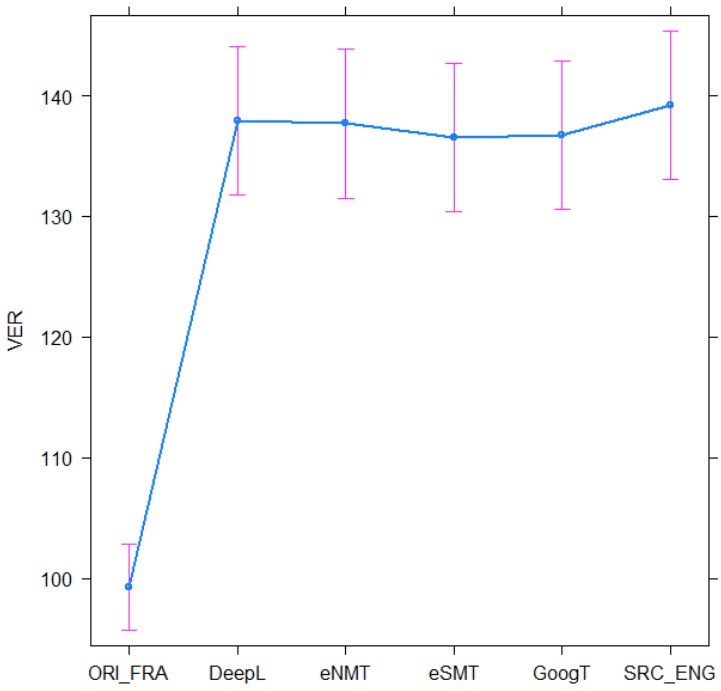


Figure 7: ANOVA analysis of the frequency of verbs (VER)

French, which is well-known to rely heavily on nominalizations, uses fewer verbs than English (compare *dans mon enfance*, ... and *when I was a child*, …). Remarkably, the corpora of machine-translated French texts all display a higher frequency of verbs than the corpus

of original French and about the same frequency as the corpus of source English texts. This hints at a shining-through effect. An increase of verbs' frequency by some 30% to over 40% compared to original French constitutes a non-negligible 'overuse'.

Given the results presented in Figure 7, one would expect a lower use of common nouns in MT French. However, as is shown by Figures 8a/b, this is not the case: common nouns (NOM) are actually (much) more frequent in MT French than in both the English source texts and original, untranslated French. Such an overuse of common nouns requires further investigation. The frequency of proper names (NAM), however, shows a highly significant decrease between the English source texts and French MT texts. Probably this is part of the explanation: proper names are somehow 'turned into' common nouns in MT. More research is clearly needed to investigate this issue.

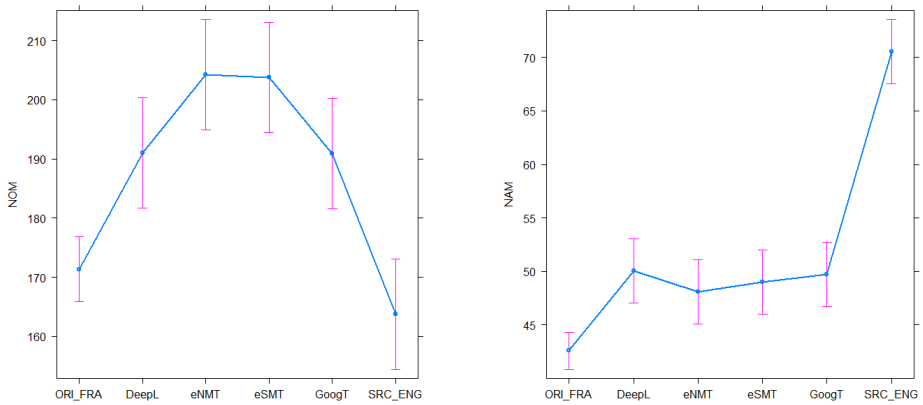


Figure 8: a/b. ANOVA analyses of the frequency of common (NOM) and proper (NAM) nouns (NOM)

5. Conclusion

In this article, we investigated the existence of machine translationese in English to French machine translations. Using the methodology and statistical techniques from corpus-based translation studies, a corpus of British English press texts was translated into French, using four different machine translation systems, and compared to French original, untranslated texts. After automatically preprocessing all texts, 22 language-independent features were extracted and subsequently the entire data matrix was analyzed with Principal Component Analysis.

This analysis revealed a distinction between original and machine-translated French, mainly due to five language-independent features: average sentence length and four features pertaining to formulaicity as expressed by combinations of three or four words or part-of-

speech combinations (ngram features). This was further explored by analysing ANOVA tests that were carried out for all features in the different settings.

Regarding sentence length, we observe a similar tendency as with human translation from English to French, namely an increase of around $20 - 25\%$. When considering the top-100 combinations of three or four words or part-of-speech tags, significant differences are found between original and machine-translated French. Similar to what was found in previous studies, the machine-translated texts thus tend to rely much more on the same word combinations, a phenomenon referred to as the "algorithmic bias" (Van Massenhove et al. 2019). Moreover, because MT systems are trained on huge amounts of human-translated parallel data this is also in line with the normalization translation universal (Baker 1993).

The ANOVA analyses also uncovered machine translationese for measures of lexical diversity. The type-token ratios of the original French are more elevated than the ones present in the machine-translated texts, corroborating previous research which found that machine translations are less lexically diverse (Toral 2019 and Vanmassenhove et al. 2019). Overall, original French exhibits more hapax, dis and tris legomena than machine-translated French, which also hints at a difference in lexical diversity. Especially the hapax legomena ANOVA yields differences among the different MT engines, suggesting that the SMT engine is closest to original French. However, upon closer inspection we discovered that this engine just leaves many words untranslated. The same analysis revealed that all MT engines also produce nonsense words, especially the eNMT engine, which is something to be avoided as this can hamper reading comprehension (Macken et al. 2019). When considering all these features, the eNMT system comes out as the one exhibiting most machine translationese and Google Translate as the one exhibiting the least.

This study has some limitations in that it only focused on original versus machine-translated French in one genre, namely press texts, and in that all features were calculated with the help of automatic preprocessing, which is not necessarily 100% accurate. Nevertheless, within a principled approach to MT tools, by both professionals and translation students who need to acquire MT literacy in order to work with the machine, such results are interesting as they provide information on what should be focused on during the post-editing process. In the case of full PE, where high quality is expected, in the light of our results, EN-FR MT output should be checked for the five linguistic features showing significant differences between machine-translated French and original French. In combination with the language-specific features investigated in Loock (2018, 2020), these independent features can provide a check-list for post-editors (e.g. reduce length of sentences), in order to try and reach linguistic homogenization with the original language, the holy grail of any translator trying to meet the invisibility demands of the high-quality end of the market.

Notes

[1] Human evaluation also has its methodological shortcomings; see Läubli et al. (2020) for an interesting discussion of which aspects should make up a human evaluation of MT output.

[2] Note that not all CBTS case studies consider the differences between original and (human) translated language to lead to translationese, a negative term suggesting that translations should be improved. Quite a number of studies actually interpret the differences as being the result of the natural influence of translation universals (simplification, normalization, explicitation, levelling out), originally defined in Baker (1993) but widely criticized since, leading rather to a "third code" for translated texts, a term that does not imply any value judgment (Gellerstam 2005, 202).

[3] https://www.deepl.com/translator

[4] www.linguee.com

[5] https://translate.google.com/

[6] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation$_e$n

[7] We would like to thank the European Commission's Directorate-General for Translation for giving us access to eTranslation.

[8] Many translation agencies often provide tables with expected expansion rates, and the one for EN-FR translation mostly amounts to $20-25\%$, see for example https://www.versioninternationale.com/details-taux+de+foisonnement+en+traduction+-+anglais+francais+allemand-395.html

[9] Please note that "least frequent" should be taken with a grain of salt as all ngram analyses are based on the top-100 most frequent ngrams.

[10] The titles of the two texts are "Zombies: A Cultural History review – a grave injustice" (Culture domain) and "Plan your own Grand Tour of Namibia - our expert's ultimate itinerary" (Travel domain).

References

Baayen, Rolf Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Babych, Bogdan. 2014. "Automated MT Evaluation Metrics and their Limitations." *TradumàTica: Tecnologies de la Traducció* 12: 464-470.

Baker, Mona. 1993. "Corpus Linguistics and Translation studies: Implications and Applications". In *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 223-250. Amsterdam and Philadelphia: John Benjamins.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. "Neural versus Phrase-based Machine Translation Quality: A Case Study." In *Proceedings of Con-*

ference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, United States, 1-5 November 2016, 257-267. http://www.aclweb.org/anthology/D16-1000 (consulted 25.09.2020).

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. "Findings of the 2015 Workshop on Statistical Machine Translation." In *Proceedings of the 10th Workshop on Statistical Machine Translation*, Lisbon, Portugal, 17-18 September 2015, 1-46. http://www.statmt.org/wmt15/pdf/WMT01.pdf (consulted 25.09.2020).

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. "Findings of the 2016 Conference on Machine Translation." In *Proceedings of the 1st conference on machine translation*, Berlin, Germany, August 2016, 131-198. https://www.aclweb.org/anthology/W16-2301/ (consulted 25.09.2020).

Bowker, Lynne, and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald Publishing.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. "Is Neural Machine Translation the New State of the Art?" *The Prague Bulletin of Mathematical Linguistics* 108: 109-120.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio-Valerio Miceli Barone, and Maria Gialama, 2017b. "A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators." In *Proceedings of the Machine Translation Summit XVI*, Nagoya, Japan, 18-22 September 2017, Vol. 1, 116-131. http://aamt.info/app-def/S-102/mtsummit/2017/conference-proceedings/ (consulted 25.09.2020).

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. "Translationese and Post-editese : how Comparable is Comparable Quality?." *Linguistica Anteverpiensia New Series – Themes in Translation Studies* 16: 89-103.

Daems, Joke, and Lieve Macken. 2019. "Interactive Adaptive SMT vs. Interactive Adaptive NMT: A User Experience Evaluation." *Machine Translation* 33(1): 117-134.

De Sutter, Gert, Marie-Aude Lefer and Isabelle Delaere . eds. 2017. *Empirical Translation Studies: New Theoretical and Methodological Traditions*. Berlin: Mouton de Gruyter.

De Sutter, Gert, Bert Cappelle, Orphée De Clercq, Rudy Loock, and Koen Plevoets. 2017. "Towards a Corpus-based, Statistical Approach of Translation Quality: Measuring and Visualizing Linguistic Deviance in Student Translations." *Linguistica Anteverpiensia New*

Series – Themes in Translation Studies 16: 25-39.

Evert, Stefan, and Stella Neumann. 2017. "The Impact of Translation Direction on Characteristics of Translated Texts. A Multivariate Analysis for English and German." In *Empirical Translation Studies. New Theoretical and Methodological Traditions*, ed. by Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 47-80. Berlin: Mouton de Gruyter.

Forcada, Mikel L. 2017. "Making Sense of Neural Machine Translation." *Translation Spaces* 6(2): 291-309.

Gellerstam, Martin. 1986. "Translationese in Swedish Novels Translated from English." In *Translation Studies in Scandinavia*, ed. by Lars Wollin, and Hans Lindquist, 88-95. Lund: CWK Gleerup.

Gellerstam, Martin. 2005. "Fingerprints in Translation". In *In and Out of English: For Better, For Worse?*, ed. by Gunilla Anderman, and Margaret Rogers, 201-213. Clevedon: Multilingual Matters.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Junczys-Dowmunt Marcin, Huang Xuedong, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Rengian Luo, Aruk Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. "Achieving Human Parity on Automatic Chinese to English News Translation." https://arxiv.org/abs/1803.05567 (consulted 25.09.2020).

Isabelle, Pierre, Colin Cherry, and George Foster. 2017. "A Challenge Set Approach to Evaluating Machine Translation." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 7-11 September 2017, 2486-2496. https://www.aclweb.org/anthology/D17-1263/ (consulted 25.09.2020).

Jenset, Gard, and Barbara McGillivray. 2012. "Multivariate Analyses of Affix Productivity in Translated English." In *Quantitative Methods in Corpus-Based Translation Studies*, ed. by Michael P. Oakes, and Meng Ji, 301-324. Amsterdam and Philadelphia: John Benjamins.

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*, Ottawa, Canada, 10-12 July 2009, 81-88.

Lapshinova-Koltunski, Ekaterina. 2013. "VARTRA: A Comparable Corpus for Analysis of Translation Variation." Paper presented at the 6th Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, August 2013.

Lapshinova-Koltunski, Ekaterina. 2015. "Variation in Translation: Evidence from Corpora." In *New directions in Corpus-based Translation Studies*, ed. by Claudio Fantinuoli, and Federico Zanettin, 93-114. Berlin: Language Science Press.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. "A Set of Recommendations for Assessing Human-Machine Parity in Language Translation." *Journal of Artificial Intelligence Research* 67: 653-672.

Laviosa, Sara. 2002. *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam and New York: Rodopi/Leiden : Brill.

Loock, Rudy 2018. "Traduction automatique et usage linguistique: une analyse de traductions anglais-français réunies en corpus." *Meta: le journal de traducteurs/Meta: Translators' Journal* 63(3): 785-805.

Loock, Rudy. 2019. "Parce que 'grammaticalement correct' ne suffit pas: le respect de l'usage grammatical en langue cible." In *La formation grammaticale du traducteur: enjeux didactiques et traductologiques*, ed. by Michel Berré, Béatrice Costa, Adrien Kefer, Céline Letawe, Hedwig Reyter, and Gudrun Vanderbauwhede, 179-194. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Loock, Rudy. 2020. "No More Rage Against the Machine: How the Corpus-based Identification of Machine-translationese can Lead to Student Empowerment." *The Journal of Specialised Translation* 34: 150-170.

Macken, Lieve, Laura Van Brussel, and Joke Daems. 2019. "NMT's Wonderland where People Turn into Rabbits. A Study on the Comprehensibility of Newly Invented Words in NMT Output." *Computational Linguistics in the Netherlands Journal* 9: 67-80.

Macken, Lieve, Daniel Prou, and Arda Tezcan. 2020. "Quantifying the Effect of Machine Translation in a High-quality Human Translation Production Process." *Informatics* 7(12). http://hdl.handle.net/1854/LU-8660184

Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty. eds. 2018. *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer.

Olohan, Maeve 2004. *Introducing Corpora in Translation Studies*. London and New York: Routledge.

Papineni, Kishore, Salim Roukos, Todd Ward, and Zhu Wei-Jing. 2002. "Bleu: a Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, United States, 7-12 July 2002, 311-318. https://dl.acm.org/citation.cfm?doid=1073083.1073135 (consulted 25.09.2020).

Rossi, Caroline, and Jean-Pierre Chevrot. 2019. "Uses and Perceptions of Machine Translation at the European Commission." *The Journal of Specialised Translation* 31: 177-200.

Tezcan, Arda, Joke Daems, and Lieve Macken. 2019. "When a 'Sport' is a Person and Other Issues for NMT of Novels.. In *Proceedings of the Qualities of Literary Machine Translation*, Dublin, Ireland, August 2019, 40-49. https://www.aclweb.org/anthology/W19-7306/ (consulted 25.09.2020).

Toral, Antonio. 2019. "Post-editese: An Exacerbated Translationese." In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, Dublin, Ireland, August 2019, 273-281. https://www.aclweb.org/anthology/W19-6627/ (consulted 25.09.2020).

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. "Attaining the Unattainable?

Reassessing Claims of Human Parity in Neural Machine Translation." Paper presented at the 3rd conference on machine translation, Brussels, Belgium, October 2018, 113-123. https://www.aclweb.org/anthology/W18-6312/ (consulted 25.09.2020).

Van Brussel, Laura, Arda Tezcan, and Lieve Macken. 2018. "A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7-12 May 2018, Miyazaki, Japan, 2018, 3799-3804. https://www.aclweb.org/anthology/L18-1600.pdf (consulted 25.09.2020).

Van de Kauter, Marjan, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Veronique Hoste. 2013. "LeTs Preprocess: The Multilingual LT3 Linguistic Preprocessing Toolkit." *Computational Linguistics in the Netherlands Journal* 3: 103-120.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. "Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation." In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, Dublin, Ireland, August 2019, 222-232. https://www.aclweb.org/anthology/W19-6622/ (consulted 25.09.2020).

Yamada, Masaru. 2019. "The Impact of Google Neural Machine Translation on Post-editing by Student Translators." *The Journal of Specialised Translation* 31: 87-106.

A Corpus-based Approach to Profiling Translation Quality: Measuring and Visualizing Acceptability of Student Translations

Yanmeng Liu

Address: School of Languages and Cultures, University of Sydney, Australia
E-mail: yliu6746@uni.sydney.edu.au
Correspondence: Yanmeng Liu
Citation: Liu, Yanmeng. 2021. "A Corpus-based Approach to Profiling Translation Quality: Measuring and Visualizing Acceptability of Student Translations." *Translation Quarterly* 101: 47-66.

Abstract

In contrast with the rapid development of translation studies as a whole, the primary purpose of this paper is to offer an innovative approach to the development of translation quality evaluation instruments using corpus data and appropriate statistical methods. Translation assessment lags behind due to its subjectivity, resulting in shifting standards of assessment and difficulty in practical operations. This article presents a corpus-based approach to profiling and assessing translation quality, more particularly translation acceptability, by comparing linguistic features in translated and original English texts with the help of machine learning methods. This article aims to give concrete statistical evidence, instead of a simple binary criterion of 'good or bad' in order to provide a more objective assessment method. To that end, the present study constructs corpus-based profiles of translated and original texts to specify the vague concept of translation quality first. Parallel Corpus of Chinese EFL Learners (PACCEL) and Lancaster-Oslo/Bergen Corpus (LOB) were selected to provide Chinese to English translations and original English texts to study. Moreover, the linguistic features were extracted for multi-level profiling in terms of lexical, syntax, and grammatical levels. Then, a multi-level comparison was conducted after testing the distinctiveness of proposed rating scales via Factor Analysis (FA) in SPSS. Finally, Decision Tree (DT) and Kernel Principal Component Analysis (KPCA) were employed to verify assessment efficiency and display the statistical results in a more straightforward manner. The results indicate that adopting a corpus-based profiling approach in translation quality analysis provides a better representation of the com-

plexities of the assessment process.

1. Introduction

Thanks to more cultural, social, economic, and political interactions among countries and regions, translation has developed and flourished rapidly, enabling people speaking different languages to communicate without language barriers. Even though translation studies have enjoyed rapid development since the 1970s, the critical research question concerning translation quality assessment (TQA) represents an important but underexplored area of research in the field (Hatim and Mason 2005), while the vast demand and prosperous development of translation make the issue of TQA an even more urgent need to be settled.

This relative neglect is primarily due to an absence of objective standards and replicable measuring methods. TQA's nature of subjectivity leads to shifting boundaries for its quality evaluation. The assessment varies as people hold different views of translation itself, which leads to different concepts of translation quality (House 2009). This largely subjective approach has long been criticized, and assessment methods based on more objective and empirical studies are called for by many scholars (Bassnett-McGuire 1991; House 2009). Another barrier causing the slow development of TQA is the low replicability of measuring methods. Previous studies have provided us with progressive understandings about translation quality, and addressed the problem of TQA from different perspectives, but the criteria are ill-defined (Hajmohammadi 2009), resulting in poor operation in practice.

With the advancement of computer technology, the novel paradigm of corpus-based translation studies (CTS) has great potential to narrow the gap by removing a great deal of subjectivity and improving the consistency of evaluation operation. Corpora enable scholars to deal with large amounts of language data and to access and retrieve empirical evidence for translation studies (Baker 1993; Laviosa 2002). From this perspective, the abstract concept of translation quality can be specified with linguistic features, and TQA can be transferred into a statistical comparison of these features. Meanwhile, advanced corpus tools and technologies provide scholars with the capability to operationalize theoretical approaches and verify hypotheses. As Ji and Oakes (2019) put it, with CTS, scholars have been moving from purely descriptive, micro-analyses of short texts to the possibility to statistically query millions of words, as a principled way to achieve representativeness and objectiveness.

The aim of the present study is to explore a more objective and replicable approach to translation quality evaluation with the help of comparable corpora, and the scope is to assess the quality of Chinese-English translations by students. My study showed that TQA could benefit from the proposed corpus-based approach in several ways: corpora can be used to profile translation quality in a more objective manner, as they analyze translations as products instead of personal understandings of them. Besides, corpus-based assessment criteria are systematic

and can compare translations on multiple levels. Finally, the thorny issue of TQA is transferred into the comparison of linguistic features extracted from corpora, which is more straightforward for both evaluators and translators to interpret. The scope of the present study is limited to a language pair of English/Chinese, described as "genetically distinct major languages in the world" (Xiao and Hu 2015, 2). The stark differences between these two distanced languages inevitably influence each other in the process of translation, which leads to distinctive features in translation results for assessment.

This article is organized as follows. After giving a concise overview of previous work on corpus-based translation evaluation in section 2, I present new concepts underlying the statistical approach in section 3. Section 4 is devoted to introducing the experiment design of the case study aimed at answering three research questions. Section 5 reports and discusses the experiment's results. The final section summarizes the main research findings and discusses the implications for future research on the evaluation of translation quality.

2. Overview of previous work

Literature shows that researchers and scholars started to sense the value of corpora in translation studies for assessing translation quality and tried to apply corpora in practice. However, the potential of corpora in translation quality assessment is far from fully explored. Generally, two types of research approaches have been explored by previous scholars. Some scholars, like Bowker (2001), Bowker and Pearson (2002), Hassani (2011), and Jiménez-Crespo (2011), believed that assessment is human activity in nature, and they treated corpora as a secondary assessment tool and subordinated references for human evaluators judging the translation quality. Some experts (for example, Rabadán et al. 2009; Loock 2017; Rojo 2018; De Sutter et al. 2017) relied heavily on the linguistic data extracted out of corpora and regarded corpora as a primary assessment tool to thoroughly implement a purely data-driven evaluation of translation quality.

For the former approach, Bowker (2001), Bowker and Pearson (2002), Hassani (2011), and Jiménez-Crespo (2011) shared the idea that corpora containing authentic texts and target language knowledge can assist evaluators in making decisions on whether the tested translation is suitable or not. In particular, Bowker (2001) and Bowker and Pearson (2002) believed that as the key in translation evaluation, trainers should possess a good knowledge of source texts, an excellent mastery of the target texts, and knowledge of the subject field. However, trainers cannot be the encyclopedia in reality, so Bowker's evaluation model concentrated on the enhancement of subject field knowledge and the target language of trainers. She compiled three types of corpora for a case study, including Quality Corpus, Quantity Corpus, and Inappropriate Corpus. Quality Corpus consisted of authentic texts written by subject field experts that provide a good explanation of the subject matter. Quantity Corpus contained a larger and

more representative sample of the language for particular purposes in question, which allows translator trainers to verify the appropriateness of the terminology. Inappropriate Corpus were unsuitable parallel texts, which helped trainers to understand why students came up with some inappropriate equivalents. Bowker's idea characterizes corpus use as a secondary tool to assist human judgement.

Hassani (2011) agreed with Bowker on evaluators' superior role in translation assessment. The difference is that Hassani set up his experiment in a professional context. His research was based on the number of predicted errors by professional translators and the number of errors detected by evaluators. The discrepancy in error count provided evidence and empirical results to prove the efficiency of corpora in translation evaluation and quality improvement. Corpus of Contemporary American English (COCA) containing authentic language expressions, useful information on style and language changes over time, and behavior of a given word, facilitated evaluators in the process of marking errors and giving feedback for translations in a test. Another scholar who tried to introduce corpus studies into translation quality assessment was Jiménez-Crespo (2011). In his understanding, the core issue was to classify translation errors, so a comparable monolingual corpus of original and translated Spanish corporate websites was used to improve error typology in his research. To achieve systematic error typology, Jiménez-Crespo first developed a genre-based description to observe the pragmatic, functional, and textual differences as potential errors. Then, corpora were used to provide quantitative verification of the mentioned errors. Finally, a case study was carried out to examine whether the proposed typology could provide an improved error-based analysis. Jiménez-Crespo categorized errors into three kinds, namely, error related to target language, pragmatic and functional errors, and localization errors. Furthermore, the evaluation was based on the accumulative number of different errors.

Bowker, Hassani, and Jiménez-Crespo shared the approach of treating corpora as auxiliary tools for evaluators to identify translation errors or inappropriate expressions. This idea is regularly adopted in the literature (see, for example, Kussmaul 1995; Delisle 2005; Collombat 2009; Dunne 2009), and a large number of corpus-based language resources are treated like a treasure for evaluators to implement the assessment. However, it is quite controversial that translation quality assessment calls for an objective assessment while insisting that translation quality assessment largely depends on human judgment. Besides, it is not plausible to state that translation is simply error counting, because translation evaluation is not about picking up faults in translation. Finally, linguistic data has great potential to dig deeper to reflect various language patterns, rather than being limited to provide examples for the assessment.

The second approach treats corpora as the primary assessment tool to implement a pure data-driven evaluation of translation quality. Seminal research started by uncovering differences between translated and original languages from the perspective of a specific linguistic feature, for example, passive constructions in translated and original Chinese (Xiao, Mcenery,

and Qian 2006), and the frequency of phrasal verbs in original English texts, English translations from Romance languages and Germanic languages (Loock 2017). These studies focused on investigating the particular singular type of linguistic feature in translated texts, which was not enough to represent the complexity of the overall translation quality. Therefore, multiple specific linguistic deviances were further applied to the mystery of translation quality assessment by several scholars.

In the studies of Rabadán, Labrador, and Ramon (2009), quantifiers, modification of nouns, and expression of the past time were selected as the object to evaluate translation quality, and their frequencies in both comparable and parallel corpora were discussed in both translated and original Spanish. Three scholars typified frequency discrepancies of mentioned linguistic features into three translation universal hypotheses: simplification (Baker 1993), law of interference (Toury 1995; Mauranen 2004), and the unique items hypothesis (Tirkkonen-Condit 2002). It was concluded that the quality of translations was in line with the similarity of grammatical structure between native and translated texts. Likewise, Loock (2017) used derived adverbs and existential constructions to measure the under-/over-representation of these two linguistic features in the translated texts, and found a correlation between observed intra-language differences and the overall quality of translations. Rojo (2018) correlated phraseological competence to the overall quality of translations and examined the translation of phraseological units in museum texts. The results showed that the texts that achieve the best result in the phraseological assessment also generally do the same in the overall assessment, and "the correlation between phraseological quality and overall quality does occur" (Rojo 2018, 13). De Sutter and his peers further enlarged the number of linguistic features under consideration when evaluating the quality of translations (De Sutter et al. 2017). They extracted over 20 language-dependent and language-independent features of translations carried out by students and professionals. A student translation corpus, a professional translation corpus, and a professional writer corpus were used for comparison. The researchers argued that it is "not very plausible to suggest that the language used by professional translators is of lower quality than that of other professional writers" (De Sutter et al. 2017, 28), so professional translations and professional writings were defined as good quality in comparison. After multi-variation analysis, ANOVA was adopted to test the significant difference between students and professionals in each linguistic feature. The overall result showed that the proposed statistical approach is feasible for translation quality assessment.

Research taking the approach of corpora as a primary assessment tool is a step toward corpus-based translation quality assessment. Compared with regarding corpora as having a subordinate role in the assessment, this approach makes full use of linguistic data extracted from corpora, in an attempt to eliminate human bias and to provide objective and empirical evidence for assessing translation quality. Furthermore, when dealing with large quantity and multi-dimension statistics, scholars start to pay attention to more advanced computer tools that

may facilitate new investigations in the field of TQA.

However, the corpus-based translation quality assessment also poses challenges for researchers. One major issue is the indicator selection. In the statistical evaluation of translations, the evaluation indicators should be identified carefully rather than picked up randomly. Unfortunately, in the research mentioned above, most of the indicators are selected, if not randomly, based on the author's personal experience. Indicators selected in this way are not systematically organized and cannot represent the overall quality effectively. This is the main reason why many of the indicators are not able to identify the underlying linguistic distinctions between good and bad translations in De Sutter's case studies (2017). Another challenge is to handle multi-variation comparison. It is significant progress to have more indicators in the translation quality assessment as the translation quality is such a complex problem to deal with, but the challenge lies in how to unify the overall result of multiple variables. This shortcoming is exposed in the research by Rabadán, Labrador, and Ramon (2009) and Loock (2017), which show that researchers can only compare the statistics of each parameter at a time, but not comprehensively analyze the statistics of the parameters as a whole.

Therefore, for the present research, the author proposes that the assessment parameters should be well-organized and systematic under the guidance of corpus linguistics, and the statistics of all parameters should be compared between translated and original texts comprehensively, with the help of advanced computer technologies.

3. Proposed concepts

In the present study, the quality of translation is evaluated with a focus on translation acceptability in the target language and culture. Concepts of acceptability distance, corpus profiling, and reader-oriented assessment are introduced to guide the research practices of TQA. Under the guidance of these three concepts, rating scales of the proposed approach are sorted and verified, translations are statistically compared, and finally, the assessment results are visualized for direct observation.

Toury (1995) is the first scholar who uses the term 'acceptability'. It refers to a translator's respect of rules and conventions of the target culture. Therefore, the judgment of acceptability is not a binary decision of right or wrong. Following Toury's idea, the present study regards acceptability as a particular point along a continuous line, with 'totally unacceptable' and 'perfectly acceptable' as the two ends of the spectrum. The tested translation should be positioned on a particular point of a consistent line with complete irrelevance at one end, and perfect acceptability at the other, as shown below. At a certain point on this line, the tested translation would have a certain distance to the perfectly acceptable expression in the target culture, which is defined as the acceptability distance of the translation. The concept of perfect acceptability is represented by a large amount of original target language data, which is a collection of texts

that the target audience can accept well.

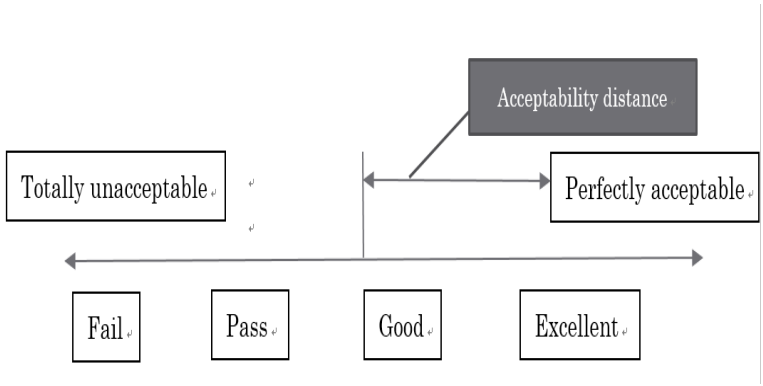


Figure 1: Diagram of acceptability distance

Worthy of mention is that the acceptability distance is a concept that helps to understand and specify the abstract idea of translation acceptability. It is not an actual variable that can be measured precisely. Furthermore, in this paper, translations with different qualities and original texts can be listed from left to right accordingly. The original texts are supposed to be located close to 'perfect acceptability' because original texts are more acceptable to target readers. While the Fail translation would be situated closer to 'totally unacceptable'. Different levels of translations and original texts are not isolated, and they are consistent with their neighboring level texts. The neighboring levels of translations do not have a clear-cut dividing line. The distinctiveness between translation and original target texts decides the acceptability distance, namely, the position of the translation along the continuous line in the diagram. Thus, the positions of texts along the line would indicate their acceptability in a more straightforward manner.

The second concept introduced is corpus-based profiling. Profiling is originally defined as "the activity of collecting information about someone, especially a criminal, to describe them" (Cambridge English Dictionary). In the definition, "a criminal" can be anyone, but with specific information or feature description of the criminal, it would be much easier to identify the criminal out of the general public. Similarly, in the context of this paper, it means that each translation or text has a statistical expression to represent itself, and the statistics are extracted from corresponding corpora. In this paper, the statistical information of each text is grouped into three levels: the lexical level, syntax level, and grammatical level. In other words, these linguistic statistics shape the features of each translation or text. For the case study, assessing Chinese to English translations by EFL students in China, this paper selected 12 indicators to profile the tested texts. That is to say, each tested text is an entity with 12 vectors or variables. The selection of 12 indicators is based on previous studies (江進林, 2013, 許家金、徐秀玲, 2016; De Sutter et al. 2017; Loock 2017; Kunilovskaya and Lapshinova-Koltunski 2019). To

be specific, under each level of parameters, detailed indicators are listed in Table 1.

Table 1: Parameters for corpus-based profiling

Corpus-based profiling	Lexical level	Type (TYPE)
		Token (TOKEN)
		Type-token ratio (TTR)
		Average word length (AWL)
		Frequently-used word ratio (FUW)
	Syntax level	Average sentence length (ASL)
		Sentence with a complex structure (SCS)
	Grammatical level	Pronouns (PN)
		Conjunctions (CONJ)
		Prepositions (PREP)
		Determiners (DTM)
		Modal words (MD)

The third concept is reader-oriented assessment. Who defines to what extent a translation is acceptable? The ideal situation to judge translation acceptability would be undertaken by the target audience or professional and experienced human raters. However, it is time-consuming and costly in practice. The machine learning techniques are an alternative way of solving the problem, as this approach is more sophisticated at learning the regularities, so as to be more accurate at approximating the human results (François and Miltsakaki 2012). Under the concept of human-oriented assessment, this study uses human raters to assess the overall quality of translations first and compares corpus-based profiles to train and test machine learning models, which will be able to predict human evaluation results afterwards, as shown in Figure 2.

When applying machine learning to translation quality assessment, the definition of to what extent the translation is acceptable is so crucial that it would influence the follow-up construction of the assessment model. The study of Petersen and Ostendorf (2009) concludes that human labelers do less well in some studies because different groups may have different evaluation results for the same text. The features of labelers are worthy of study because the human labelers actually represent readers who share similar features with them, and the use of machine learning can be a means of tuning the assessment models to the needs of a particular group of readers. This means that when labeling, human raters shall represent a particular group of readers, e.g., translation trainers, professional practitioners, etc.

With these three concepts, the assessment of translation quality turns into a judgment of acceptability distances via comparing corpus profiles of translations and original texts. The assumption underlying this research is that the statistics in corpus profiles of texts of simi-

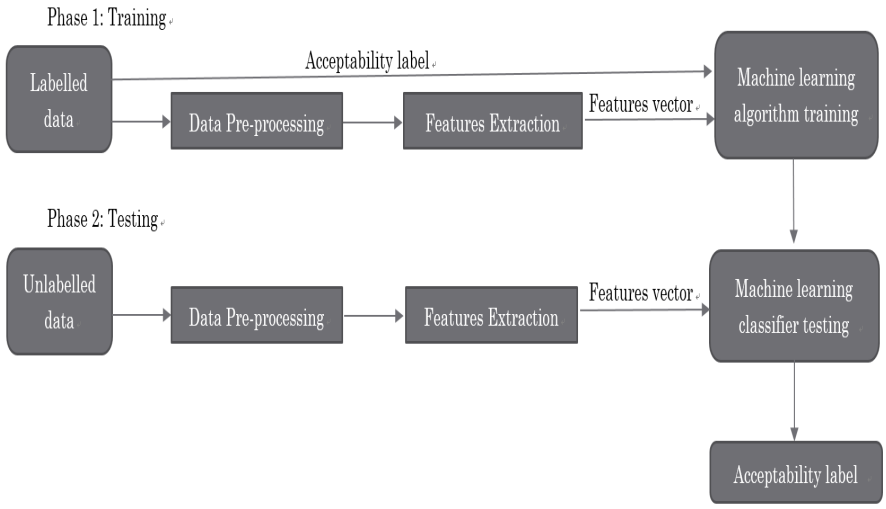


Figure 2: Machine learning for translation acceptability assessment

lar quality are relatively homogeneous to its kind, and the native text profiles would form a standard, implying what translations with good acceptability should be like. The underlying hypothesis of this study is that the homogeneity of original text profiles contains perfect acceptability concerning the rules and conventions in target language and culture, and better translation profiles shall be closer to this homogeneity but not completely overlapping with original text profiles. Represented by their corpus-based profiles, different levels of translations can be assessed by means of acceptability distance with the approach of machine learning.

4. Experimental design

The experiment is designed to answer the following research questions:

1.How can the translation acceptability be statistically assessed with corpus-based data?

2.Which predefined assessment indicators are strong enough to distinguish translation acceptability at different levels?

3.To what extent does the machine learning model accurately predict the human rating results?

To answer these research questions, the assessment comprises three steps: (1) corpora selection and initial data collection, (2) relevance and accuracy verification of assessment indicators, (3) assessment result visualization.

(1)Corpora selection and initial data collection

In order to have representative data for corpus profiling, corpora were carefully selected. The translation samples were sourced from the Parallel Corpus of Chinese EFL Learners (PACCEL), a two-million-word corpus of Chinese-English translations. This corpus was divided into two parts: Parallel Corpus of Chinese EFL Learners—Written translation (PACCEL-W) (1.6 million words) and Parallel Corpus of Chinese EFL Learners—Spoken interpreting (PACCEL-S) (0.5 million words) (文秋芳、王金銓, 2008). And each student translation was scored out of 100 by human raters and was further classified into four levels: Excellent (100-80), Good (79-70), Pass (69-60), and Fail (less than 59). The original score of each translation was the average mark rated by two professional translation trainers. When the rating results of the same translation fell into different quality levels, the two trainers were requested to rate again till they fell into the same level (Ibid. 2008,12). The original score marked by translation trainers would be treated as labels for machine learning algorithm training and as a reference to verify the proposed assessment method. And the labelling of the student translation is based on the judgment of professional translation trainers. As discussed earlier, the machine learning model would be built upon acceptability evaluation from the perspective of translation trainers. That is to say, machine learning would tune itself to approximate translation trainers' evaluation results.

The study selected student translation materials as experiment samples for several reasons. Firstly, student translation could provide diverse samples with different degrees of acceptability to study. As translation learners, students were on their way to becoming competent translators able to produce qualified translation with good acceptability. The student translations varied widely, ranging from fail ones to excellent ones, which were suitable to train and test the proposed method. Secondly, student translations provided a large number of translation samples for the same original text, which would be advantageous for more precise quality estimation, because the interference caused by differences related to the source texts was avoided. Otherwise, the acceptability assessment of translations for different original texts would disturb the construction of the assessment model. Thirdly, the materials were translated by students sharing similar demographic characteristics, which made the translations they produced comparable.

The selected Chinese to English translation was about a celebrity's opinion on movie industry development. Furthermore, in order to eliminate the influence caused by an uneven distribution of data, this research selected 80 samples randomly, with 20 samples for each level of translation. And the materials of the comparable or reference corpus were selected from the Lancaster-Oslo/Bergen Corpus (LOB) (Johansson, Leech, and Goodluck 1978). Considering the interference of register and genre, this research selected subcategories of Press: Reportage, Press: editorial, Press: reviews, Biography, General fiction in the LOB. And these texts were randomly grouped into 20 groups as 20 samples for original English. As described in the previous section, this study would profile 100 translation samples and original texts with 12

types of linguistic features extracted from corpora in order to statistically assess the translation acceptability. The corpus tool of LancsBox 4.0 was adopted to extract these linguistic features (Brezina et al. 2018). Table 2 shows the overall profiles of five categories of texts.

Table 2: Overall profiles of five categories of texts

Parameters	Indicator	Original(O)	Excellent(E)	Good(G)	Pass(P)	Fail(F)
Lexical level	TOKEN	391247	10144	15407	26180	10147
	TYPE	29474	1108	1367	1866	1087
	TTR	0.075333485	0.109227129	0.088725904	0.071275783	0.107125259
	AWL	2.58213352	2.527509	2.466844	2.481005	2.399861
	FUW	15423	573	731	942	572
Syntax level	ASL	13.2427736	14.01536	13.81315	13.93453	14.16942
	SCS	4436	68	101	172	58
Grammatical level	PN	26782	253	1359	659	914
	CONJ	16102	448	687	1049	1357
	PREP	52226	1118	1652	2808	1113
	DTM	4425	801	1282	2242	918
	MD	5288	256	414	694	266

Similarly, under an umbrella of three parameters, each of the samples was profiled with 12 corpus-based linguistic statistics, which formed the initial data of the study. And the initial data would be filtered and processed through the following two steps.

(2)Relevance and accuracy verification of assessment indicators

The mentioned 12 linguistic features were initially regarded as indicators to assess translation acceptability in the proposed approach. The research employed SPSS Factor Analysis to test to what extent the indicators were targeting the same construct and to streamline these initial indicators. Indicators with high factor loadings (≥ 0.5) were kept as final assessment indicators, and final assessment indicators contributing to the same principal component (Eigenvalue ≥ 1.0) would be grouped into one parameter for the acceptability assessment. This process filtered and streamlined the predefined linguistic features, avoiding disturbance of unrelated features for machine learning later. The pre-processing of data eliminated disturbing factors, which would potentially improve the accuracy of training and testing for machine learning.

After finalizing the assessment indicators, the efficiency and representativeness of these indicators were proved through Decision Tree (DT) analysis. Supervised machine learning was adopted in this process, with sample texts having been divided randomly into two categories as training and testing data. Typically, 80% of labelled data (80 samples in the present study) would be training data, and the remaining 20% (20 samples) were testing data. The training data were used to train the algorithm via DT to classify translations of different qualities and original texts, while the testing data were used to test the accuracy of the trained algorithm, as shown in Figure 2 above. The testing accuracy demonstrated the capability of distinguishing

student translations of different levels and original texts. The higher the testing accuracy was, the more accurate the proposed method was in predicting the acceptability of translations. And the confusion matrix displayed the detailed classification of prediction by DT compared with human assessment by professional translation trainers.

(3)Visualization of the assessment result

Finally, an integrated assessment result was presented, and the numerical information gathered from these corpora was analyzed with Kernel Principle Component Analysis (KPCA), which enabled a more straightforward display of complex data. As mentioned earlier, each corpus profile contained multiple variables, also known as vectors. KPCA could convert possibly correlated variables into a set of values of non-linearly uncorrelated variables. Through analyzing the corpus profiles and linguistic behaviors of student translations, the multi-vector comparison could visualize acceptability distance between different translation and text profiles. And the visualization of acceptability assessment results referred to the acceptability of translation compared with original English texts. The visual representations would elucidate the extent to which different levels of translations approximate to original English.

5. Results

The results showed that 12 predefined indicators demonstrated high importance to contribute to three principal components, namely, Expression Diversity, Expression Accuracy, and Expression Complexity. With these 12 indicators, the machine learning tool was able to distinguish translation acceptability levels with an accuracy rate of 80%. And the visualized results indicated that Expression Diversity had the most considerable influence on the translation acceptability assessment.

5.1 The relevance of assessment indicators

The initial statistics of 100 samples were input into the SPSS Factor Analysis. Factor loadings of each initial indicator or factor indicated that the dimensions of the factors were better accounted for by the variables (Yong and Pearce 2013). Therefore, the loadings reflected the relative importance of 12 indicators. As shown in the analysis result in Table 3, all of 12 indicators obtained a factor loading beyond 0.5, and three principal components were extracted. To be specific, Component 1 included token (TOKEN), type (TYPE), token-type ratio (TTR), frequently-used words (FUW), average sentence length (ASL), modal (MD), pronoun (PRON), and conjunctions (CONJ). Component 2 consisted of preposition (PREP) and determiner (DTM). Component 3 was made up of average word length (AWL) and complex sentence (CS).

From a corpus linguistics perspective, Component 1 could be named as Expression Diversity. Token, type and token-type ratio, modal, and frequently-used words were indicators

Table 3: Factor Analysis results

Indicators	Component		
	1	2	3
TOKEN	0.959	0.194	0.102
TYPE	0.97	0.185	0.083
TTR	-0.924	0.235	-0.002
AWL	0.073	0.311	0.617
FUW	-0.837	0.391	-0.093
ASL	0.835	0.257	0.286
CS	0.108	-0.21	0.779
PREP	0.087	0.846	0.219
DTM	0.094	0.796	-0.101
MD	-0.672	0.311	-0.217
PRON	0.812	0.173	0.062
CONJ	0.927	0.189	0.106

to show the lexical richness. And average sentence length, pronoun, and conjunctions represented syntactic richness. They both explained the diversity of articulations. Component 2 was defined as Expression Accuracy. Even though prepositions and determiners had different grammatical functions in an expression, they both modified nouns in their own ways. The function of modifying could polish the expression to make it more accurate. Component 3 was identified as Expression Complexity. Average word length indicated the complexity of wording, and the complex sentence referred to sentences containing a subordinate clause or clauses, which was also the complexity of sentences.

As a result of the current stage, 12 indicators were tested to be of high relevance with the distinctiveness among five categories of texts. And these indicators could be retained as three principal components, which were further developed into the multi-dimensional comparison in the following sections.

6. Accuracy of the proposed assessment method

100 samples, each of which contains 12 vectors, were input into DT to analyze whether the evaluation results of the proposed approach match human assessment results. As a supervised learning model, DT divided a dataset with a given answer into training data (80 samples) to train the algorithm for classification and testing data (20 samples) to check the result of a trained model. The result of testing accuracy was 80%, and the confusion matrix in Table 4 showed detailed evaluation results.

Table 4: Confusion matrix for testing samples

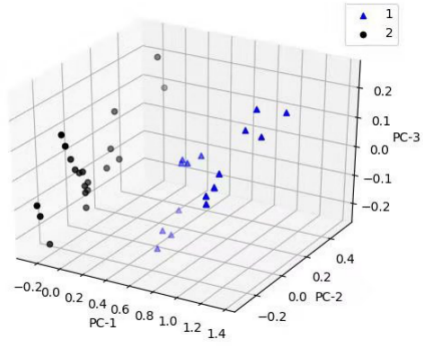
	O	E	G	P	F
O	4	0	0	0	0
E	0	4	0	0	0
G	0	0	2	1	1
P	0	0	1	3	0
F	0	0	0	1	3

The testing accuracy of 80% demonstrated that with the aforementioned 12 features as indicators, translation quality can be correctly estimated with a high accuracy. That is to say, these 12 features exhibited a significant difference among original English, Excellent, Good, Pass, and Fail translations. In Table 4, Original English was marked as O, and Excellent translation as E, Good translation as G, Pass translation as P, and Fail translation as F. Almost all the samples were located into the categories they should be, while the classification was wrong when judging the two Good translations into Pass and Fail. One Pass translation was wrongly classified into the category of Good, and one Fail translation was wrongly put under the Pass category. Generally, as indicated in Table 4, the assessment results of 100 samples by proposed approaches were mostly matched with human assessment results, and the accuracy was high, with minor errors in the evaluation.

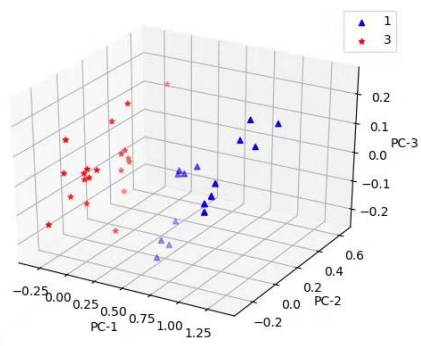
6.1 Visualization of the assessment result

Then, 100 samples were analyzed via KPCA to visualize the multi-vector comparison. As described above, each text was profiled with 12 vectors, which were extracted from corresponding corpora. However, a human cannot visualize a 12-dimensional phenomenon, so the present study adopted KPCA, which could convert a set of various numerical values into two or three principal components and display the result in a three-dimensional or two-dimensional layout. The result of KPCA was visually presented in the Figures below.

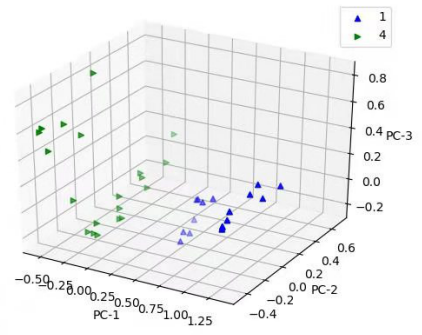
It was noted that every observable spot represented a sample text or translation. And five categories of samples were marked in five colors. Those in blue were original English texts, those in black were Excellent translations, Good translations were red, Pass translations were green, and Fail translations were yellow. The x-, y- and z-axis represented Component 1, Component 2, and Component 3. What was of great value here was the relative position of the different spots in this coordinated system. The closer two texts were, the more similar they were in terms of 12 linguistic features. This means that the closer one text was to the original English texts, the shorter its acceptability distance was, which represented to what extent the translated text was acceptable in the target language and culture of professional translation trainers.



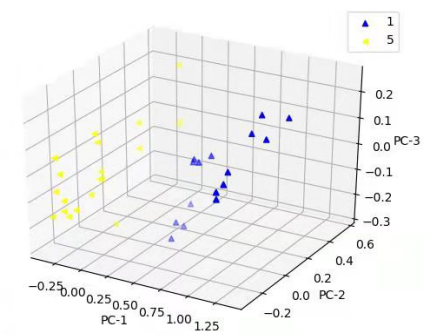
(a) Figure 3. Three-dimension plot for E & O



(b) Figure 4. Two-dimension plot for G & O



(c) Figure 5. Three-dimension plot for P & O



(d) Figure 6. Three-dimension plot for F & O

Pairwise comparisons were employed in Figure 3 to Figure 6, which showed the comparison results between Excellent translations and original English texts, Good translations, and original English texts, Pass translations and original English texts, Fail translations and original English texts in three-dimensional plots. It was observed that in general, original English texts were divided from the translations by students, as the blue spots were positioned separately from the spots in black, red, green and yellow in the Figures. The original English texts got together to form a cluster, which represented perfect acceptability. In comparison, all the student translations, be it an Excellent one or Failed one, had an acceptability distance away from the perfect cluster.

As discussed in previous sections, all the 12 indicators would be retained into principal components of Expression Diversity, Expression Accuracy, and Expression Complexity. Expression Diversity had the most considerable influence on the assessment of translation acceptability, followed by Expression Accuracy and Expression Complexity. As shown in the Figures above, student translations suffered from low acceptability mainly because of Expression Diversity, as the larger distances between student translations and original English exist in x-axis, representing the principal component of Expression Diversity. In contrast, the distances in the dimension of the y-axis or z-axis were close to the position of original English texts.

This referred to the fact that students did an excellent job in terms of Expression Accuracy and Expression Complexity.

7. Conclusion

This paper has examined the assessment of translation acceptability using corpus-based indicators combined with the machine learning approach. The concepts of acceptability distance, corpus-based profiling, and reader-oriented assessment transfer the abstract question of quality assessment into a statistical comparison of linguistics features, which facilitates the objective and feasible assessment of translation quality. The linguistic features extracted from corpora are grouped into three dimensions: Expression Diversity, Expression Accuracy, and Expression Complexity. These three dimensions specify the definition of acceptability distance, and are further displayed in more straightforward plots. The assessment accuracy of the proposed approach is 80%, and the visualization of the assessment results is able to indicate the quality of student translations through the comparison with original English texts. The proposed approach would benefit the field of TQA in several ways. Theoretically, three new concepts are introduced in this paper to sketch out the translation quality in a more objective way. Besides, the methodology introduced makes the TQA more operational in practice. The empirical study and statistical analysis of translation quality enhance the feasibility of TQA. Furthermore, the advancement of corpus tools and other computer technologies enables researchers to extend the usage of corpora in TQA and explore potentials of interdisciplinary cooperation. Last but not least, tools like machine learning improve the translation quality assessment by using data to approximate human assessment results, making it more reader-oriented. Also, they can be tuned according to different research or practical needs.

However, there are some inherent limitations in the current research. The proposed indicators cannot cover all the elements that influence the acceptability of translations. These indicators are mainly from lexical, syntactic, and grammatical perspectives, while the acceptability of translations may also be affected by other indicators, such as semantic or cultural factors. Further research is required if a comprehensive assessment of translation acceptability or translation quality at large is to be achieved. Besides, the number of samples and languages in the experiment is limited in the present study, which calls for empirical studies of a larger scale and in more language pairs to enhance the consistency of the proposed approach. Finally, this study only deploys one machine learning algorithm of Decision Tree to train and test the assessment model. Actually, there are many other options to choose from, and the comparative experiment with different computer technologies is suggested in order to explore more accurate applications of machine learning in corpus-based translation studies.

In the future, hopefully, this paper will encourage more researchers in the field to investigate the challenging issue of TQA from the perspective of corpus-based approaches. Moreover,

the interface between corpus linguistics and technological advancement may contribute to the development of TQA research and elucidate the complex process of TQA practice.

Acknowledgements

The work described in this paper was supported by Raymond Hsu Scholarship (No. SC1450), AR Davis Postgraduate Research Scholarship (No. SC3234), and FASS Research Bursary Scholarship (No. SC3422) from The University of Sydney.

References

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies: Implications and Applications." In Text and Technology: In *Honour of John Sinclair*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233-250. Amsterdam and Philadelphia: John Benjamins.

Bassnett-McGuire, Susan. 1991. *Translation Studies*. London and New York: Methuen.

Bowker, Lynne. 2001. "Towards a Methodology for a Corpus-based Approach to Translation Evaluation." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 46(2): 345-364.

Bowker, Lynne, and Jennifer Pearson. 2002. *Working with Specialized Language: a Practical Guide to Using Corpora*. London and New York: Routledge.

Brezina, Vaclav, Matthew Timperley, and Anthony McEnery. 2018. "# LancsBox v. 4. x.". Lancaster University.

Cambridge English Dictionary. https://dictionary.cambridge.org/dictionary/english/profiling. Accessed 20 November 2019.

Collombat, Isabelle. 2009. "La Didactique de L'erreur Dans L'appretissage de la Traduction." *Jostrans: The Journal of Specialized Translation* 12: 37–54.

De Sutter, Gert, Bert Cappelle, Orphee De Clercq, Rudy Loock, and Koen Plevoets. 2017. "Towards a Corpus-based, Statistical Approach to Translation Quality: Measuring and Visualizing Linguistic Deviance in Student Translations." *Linguistica Antverpiensia, New Series–Themes in Translation Studies* 16: 153-166.

Delisle, Jean. 2005. *L'Enseignement Pratique de la Traduction*. Ottawa: Presses de l'Université d'Ottawa.

Dunne, Keiran. 2009. "Assessing Software Localization: Toward a Valid Approach." In *Testing and Assessment in Translation and Interpreting Studies*, ed. by Claudia V. Angelelli, and Holly E. Jacobson, 185-222. Amsterdam and New York: John Benjamins.

François, Thomas, and Eleni Miltsakaki. 2012. "Do NLP and Machine Learning Improve Traditional Readability Formulas?." In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, Montreal*, 2012, 49-57. Montreal: Association for Computational Linguistics.

Hajmohammadi, Ali. 2009. "Translating Law." *Perspectives* 17(3): 211-212.

Hassani, Ghodrat. 2011. “A Corpus-based Evaluation Approach to Translation Improvement.” *Meta: Journal des Traducteurs/Meta: Translators’ Journal* 56(2): 351-373.

Hatim, Basil, and Ian Mason. 2005. *The Translator as Communicator*. New York: Routledge.

House, Juliane. 2009. “Quality”. In *Routledge Encyclopedia of Translation Studies*, ed. by Mona Baker, and Gabriela Saldanha, 222-225. London and New York: Routledge.

Ji, Meng, and Michael P. Oakes. 2019. *Advances in Empirical Translation Studies: Developing Translation Resources and Technologies*. Cambridge: Cambridge University Press.

Jiménez-Crespo. 2011. “A Corpus-based Error Typology: Towards a More Objective Approach to Measuring Quality in Localization.” *Perspectives* 19(4): 315-338.

Johansson, Stig, Geoffrey N. Leech, and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computer*. Oslo: Department of English, University of Oslo.

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, Varna, 2019*, 47-56. Shoumen: Incoma Ltd.

Kussmaul, Paul. 1995. *Training the Translator*. Amsterdam and Philadelphia: John Benjamins.

Laviosa, Sara. 2002. *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam and New York: Rodopi/Leiden: Brill.

Loock, Rudy. 2017. “Intra-language Differences and Translation Quality Assessment: An Exploratory Study on a Learner Corpus of Literary Translations.” *inTRAlinea Special Issue: Corpora and Literary Translation*, http://www.intralinea.org/specials/article/2258.

Mauranen, Anna. 2004. “Corpora, Universals and Interference.” In *Translation Universals: Do They Exist*, ed. by Anna Mauranen, and Pekka Kujamaki, 65-82. Amsterdam and Philadelphia: John Benjamins.

Petersen, Sarah E., and Mari Ostendorf. 2009. “A Machine Learning Approach to Reading Level Assessment.” *Computer Speech & Language* 23(1): 89-106.

Rabadán, Rosa, Belén Labrador, and Noelia Ramón. 2009. “Corpus-based Contrastive Analysis and Translation Universals: A Tool for Translation Quality Assessment English–> Spanish.” *Babel* 55(4): 303-328.

Rojo, Jorge Leiva. 2018. “Phraseology as Indicator for Translation Quality Assessment of Museum Texts: A Corpus-based Analysis.” *Cogent Arts & Humanities* 5(1), 1442116. DOI: 10.1080/23311983.2018.1442116.

Tirkkonen-Condit, Sonja. 2002. “Translationese—a Myth or an Empirical fact? A Study into the Linguistic Identifiability of Translated Language.” *Target. International Journal of Translation Studies* 14(2): 207-220.

Toury, Gideon. 1995. “The Notion of ‘Assumed Translation’-An Invitation to a New Discussion”. *Letterlijkheid, Woordelijkheid/Literality, Verbality*: 135-147.

Xiao, Richard, Tony McEnery, and Yufang Qian. 2006. "Passive Constructions in English and Chinese: A Corpus-based Contrastive Study." *Languages in Contrast* 6(1): 109-149.

Xiao, Richard, and Xianyao Hu. 2015. *Corpus-based Studies of Translational Chinese in English-Chinese Translation*. Heidelberg: Springer.

Yong, An Gie, and Sean Pearce. 2013. "A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis." *Tutorials in Quantitative Methods for Psychology* 9(2): 79-94.

江進林 (2013), "英譯漢語言質量自動量化研究", *現代外語* 1: 85-91。

文秋芳、王金銓 (2008), *《中國大學生英漢漢英口筆譯語料庫》*。北京：外語教學與研究出版社。

許家金、徐秀玲 (2016), "基於可比語料庫的翻譯英語銜接顯化研究", *外語與外語教學* 6: 94-102, 122。

Translation Universals in Legal Translation: A Corpus-based Study of Explicitation and Simplification

Francesca Luisa Seracini

Address: Università Cattolica del Sacro Cuore, Milan, Italy
E-mail: francesca.seracini@unicatt.it
Correspondence: Francesca Luisa Seracini
Citation: Seracini, Francesca Luisa. 2021. "Translation Universals in Legal Translation: A Corpus-based Study of Explicitation and Simplification." *Translation Quarterly* 101: 67-91.

Abstract

Research into legal translation has drawn attention to the potential effects that universal tendencies in translation could have on translated legal texts (Biel 2010; Prieto Ramos 2014; Pontrandolfo 2020), with particular regard to accuracy and readability. This paper takes simplification and explicitation into consideration and investigates the impact of these two tendencies on translated legislative texts. The analysis was carried out on a parallel corpus of EU legislative texts translated from English into Italian and a reference corpus of national non-translated legislation. A corpus-based approach was adopted with a view to providing quantitative data that could be indicative of a tendency to simplification and explicitation in the translated texts. A subsequent qualitative analysis carried out on the parallel corpus examined the shifts between original and translated legislative texts, with a focus on the linguistic features that point to an increased level of simplification and explicitation in the translations. The results of the analysis were compared to data obtained from the reference corpus in order to identify those elements that characterise the translated legislative texts differently from non-translated legislation. The results revealed that the translated texts tend to be more explicit than the source texts at a lexical and morphosyntactic level. Evidence of simplification was also found in the tendency to omit unnecessary repetitions, to avoid complex sentence structures and to prefer the active voice. The paper discusses the implications of simplification and explicitation in EU translated legislation: while they contribute to clarity and readability, they also reduce vagueness and limit interpretation.

1. Introduction

The present paper draws on the findings from previous corpus-based studies conducted by the author (Seracini 2019; 2020) on the EURO-CoL corpus, where research into translated EU legislation provided evidence in support of some of the theories on translation universals. The analysis in the present paper extends the research further, providing a more in-depth insight into explicitation and simplification in legal translation.

Research into the language of translated texts found that it generally presents features that deviate from the language of comparable texts originally drafted in the target language. These deviations do not depend on the level of competence of the translator (Toury 1979, 226), but, as Baker (1993, 242) observes, derive from "the very activity of translating, the need to communicate in translated utterances, [which] operates as a major constraint on translational behaviour and gives rise to patterns which are specific to translated texts". In Baker's (1993, 243) definition, these patterns are "universal features", "which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems". Chesterman (2004a; 2004b) classifies universals into two categories: "S-universals" and "T-universals". The former can be detected in the shifts between source texts and target texts; the latter can be identified by comparing translations with comparable non-translated texts.

Various scholars have questioned the idea of 'universality' of these features (e.g. Tymoczko 1998; Chesterman and Arrojo 2000; House 2008). However, as Chesterman (2004b, 11) points out, "[w]hat ultimately matters is perhaps not the universals, which we can never finally confirm anyway, but new knowledge of the patterns, and patterns of patterns, which help us to make sense of what we are looking at". In Chesterman's (2010, 46) view, "perhaps it would be more fruitful to search for less-than-universal patterns in translation profiles, under different sets of conditions, and thus make more modest claims". Awareness of T-universals, for example, can help practitioners avoid certain features that would make the translated texts sound unnatural in the target culture (Chesterman 2004b, 11). From a pedagogical point of view, the quality of the students' translations can improve if they are taught about undesirable features that are recurrent in translated texts (Chesterman 2010).

As regards Legal Translation Studies (cf. Prieto Ramos 2014), Biel (2010, 8) remarks that "[t]ranslation universals elicit a number of questions, still unanswered, concerning their potential impact on legal translation" and their effect on the "accuracy and naturalness of translations". However, as Pontrandolfo (2020, 23) observes, in legal translation, "universals have not been tested extensively, possibly due to the absence of large legal corpora that could be used as testbeds to confirm or disconfirm such patterns". The importance of further corpus-based research on universals in legal translation is emphasised by Pontrandolfo (2019, 22), who points out that "[t]he use of legal corpora effectively helps scholars to isolate descriptive features of translations that actually give insights into the complex dynamics of legal transla-

tion". As mentioned above, the present paper intends to contribute to research into translation universals in legal translation by investigating EU legislative texts translated from English into Italian. The focus of the analysis is on two tendencies that are of particular relevance for the translation of multilingual legislation, i.e. explicitation and simplification.

The legislation adopted by the European Union is made available in all the 24 official languages, as established by Council Regulation 1/58[1] and all language versions are equally authoritative. Despite the fact that the Regulation mentions 'drafting' and not 'translation', the production of the different language versions involves, in fact, a translation process from the original draft – usually in English. [2]

There are specific guidelines and common rules as regards the drafting and translation of EU legislation, which ensure that all the language versions are consistent and uniform in terms of structure and terminology. The *Joint Practical Guide*[3] (European Union 2015, 10), for example, requires that legal acts are "clear, easy to understand and unambiguous", "simple and concise, avoiding unnecessary elements", and "precise, leaving no uncertainty in the mind of the reader". "Naturalness" in translated texts is also a desired feature, as mentioned in the *DGT Translation Quality Guidelines* (European Commission Directorate-General for Translation 2015, 2).[4] As the Communication from the Commission (COM(2015) 215 final), *Better regulation for better results – An EU agenda*[5] specifies, EU laws should be drafted in such a way, that they are "correct, comprehensible, clear, and consistent" in order to enable everyone to "understan[d] their rights and obligations easily and with certainty". Clarity and readability in legislative texts are, therefore, key requirements. The institutional guidelines and established conventions, as well as the principle of equal authenticity of the different language versions impose constraints on the work of both drafters and translators (Ulrych 2014, 16-17). The present study takes these constraints into consideration when considering the results of the analysis on explicitation and simplification in the EURO-CoL corpus. Sections 1.1 and 1.2 report on previous research into the two universals of translation, with particular reference to legal translation.

1.1 Explicitation

The concept of explicitation was first introduced by Vinay and Darbelnet ([1958]1995, 342) who define it as "a stylistic translation technique which consists of making explicit in the target language what remains implicit in the source language because it is apparent from either the context or the situation". Studies investigating explicitation have developed this concept further. Séguinot (1988, 108) argues that explicitation occurs when "something which was implied or understood through presupposition in the source text is overtly expressed in the translation, or an element in the source text is given a greater importance through focus, emphasis, or lexical choice". In Baker's (1996, 180) view, explicitation is "an overall tendency to spell things out rather than leave them implicit in translation". Scholars found evidence in

favour of an explicitation tendency in a number of shifts between source texts and target texts. For example, shifts were found in the use of cohesive markers (Blum-Kulka 1986), in the higher number of linking words and changes to the sentence structure in the target texts (Séguinot 1988), in the increased frequency of optional 'that' after the verbs 'say' and 'tell' (Olohan and Baker 2000) and in the increased length of translated texts compared to their originals (Baker 1996). In Chesterman's (2004a) classification, explicitation, as well as "lengthening", i.e. the fact that translations tend to be longer than their source texts, are regarded as potential "S-universals", that is the recurrent features found in translated texts that point to differences between source texts and target texts.

Explicitation can involve either addition or specification (Faber and Hjort-Pedersen 2013). In the case of addition, "extra lexical items that either add or repeat meaningful elements" are introduced in the target text (Faber and Hjort-Pedersen 2013, 44). In the case of specification, "lexical elements that are semantically more informative than the ST lexical elements" are used in the target text, which results in the fact that meanings are added to the translations (Faber and Hjort-Pedersen 2013, 44). Therefore, while the measure of addition is quantitative, the measure for specification is qualitative.

Klaudy (1998) distinguishes between 1) "obligatory explicitation", 2) "optional explicitation", 3) "pragmatic explicitation", and 4) "translation-inherent explicitation". Obligatory explicitation refers to necessary changes at a grammatical/lexical level in the target texts, due to linguistic reasons. Instead, optional explicitation refers to changes introduced to improve the naturalness of the target language in the translated text. Pragmatic explicitation concerns the addition in the target text of information that the target audience lacks (e.g. culture-bound references). Translation-inherent explicitation is instead due to the translation process itself that sometimes leaves traces in the translated text.

The concept of explicitation as a translation-inherent feature has attracted some criticism (e.g. House 2008; Becher 2010a; 2010b). In particular, some studies have pointed out that directionality, i.e. the relation between source language and target language, affects explicitation, which would, consequently, contradict the view that explicitation could potentially be a universal feature of translation. In House's (2008, 12) words, "candidates of universality suggested for one particular translation direction need not necessarily be candidates for universality in the opposite direction". Munday's (1998) research into the English translation of the short story by Gabriel García Márquez, *Diecisiete ingleses envenenados* also provides evidence of the fact that certain systemic differences between English and Spanish, such as the omission of subject pronouns in Spanish, are responsible for the higher number of tokens in the target text. As Munday (1998, 4) observes, "[t]he comparative length of the ST and TT may depend on many variables, and seems to be an area far more complex than previously thought and worthy of careful future investigation on other texts".

As mentioned earlier, further research is needed in order to confirm or disprove claims

about translation universals. As regards explicitation in legal translation, considering that, as Hjort-Pedersen and Faber (2010, 238) observe, "[f]rom a legal point of view, adding or subtracting information in legal translation is a high-risk procedure because of the potential change of legal meaning and/or effect of the target text", more in-depth knowledge would be required in order to establish its influence on the quality of the translated texts (Krogsgaard Vesterager 2017).

Biel (2014, 100) found that explicitation correlates with the "conceptual distance between legal systems", that is, "the more distant legal systems are the higher the need to explicate" (cf. also Mauranen 2007). An indicator of explicitation in legal translation was identified in the higher frequency of linking adverbials introducing apposition, contrast and concession in legal texts translated from Spanish into English (Pontrandolfo 2020), which signals an attempt to add clarity to sense relations in the target texts. Krogsgaard Vesterager's (2017) study of explicitation in the translations of a judgment from Spanish into Danish carried out by experts/non-experts provided evidence of the fact that explicitation – in particular addition - occurred most frequently in the translation of system-bound terms and elliptical phrases. The study also demonstrated that expert translators tend to explicitate more than non-expert translators.

1.2 Simplification

As a universal feature that occurs in translated texts, simplification is defined as a tendency to simplify the language at a lexical, syntactic and textual level (Baker 1996). Since simplifying a text also means making it more explicit, the boundary between simplification and explicitation is not always clear-cut (Baker 1996, 182). In Baker's (1996, 182) words, "simplification involves making things easier for the reader (but not necessarily more explicit), but it does tend to involve also selecting an interpretation and blocking other interpretations, and in this sense it raises the level of explicitness by resolving ambiguity". In Chesterman's (2004a) classification, simplification is a potential T-universal, i.e. it is among the recurrent features that are signalled by differences between translations and comparable non-translated texts.

At a lexical level, Blum-Kulka and Levenston (1983, 119) define simplification as "the process and/or result of making do with *less* words" (emphasis in the original). In line with this definition, Laviosa's (2003, 159) study of lexical simplification demonstrates that "translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower)". The complexity of identifying simplification in a translated text is highlighted by Mauranen (2007, 40), who maintains that word combinations, and not only single words, should be considered.

With respect to legal translation, Biel (2014) points out that avoidance of repetitions – a recurrent tendency in translation (Blum-Kulka and Levenston, 1983) - and disambiguation are the two elements related to simplification that are of particular relevance. Distinguishing between ambiguity and vagueness, Engberg and Heller (2008, 148) observe that ambiguity in

legal drafting is "in communicative terms […] a rather undesirable feature", while vagueness can be "an accepted, necessary element". As Antelmi (2008, 94) remarks, "vagueness would appear to have the advantage of flexibility, instead of the disadvantage of imprecision" [my translation],[6] since it concedes some flexibility to the legal terms, enabling them to cover, over time, new concepts and situations. With particular regard to the translation of multilingual legislation, Biel (2014, 103) cautions against the risks of disambiguation since a certain degree of ambiguity or vagueness could be intentional, as multilingual law "relies on a political consensus".

In his study on the translation of the Spanish Constitutional Court's judgments into English, Pontrandolfo (2020) investigated lexical variety, lexical density and mean sentence length in a parallel corpus and found evidence of simplification in the first two indicators, but not in the third. Anselmi and Seracini's (2015) study of intralingual translation in a corpus of EU directives transposed into British law points to lower lexical density and disambiguation in the implemented legislation, thus providing evidence of the fact that the tendency to simplify language also characterises legal translation. In Seracini (2019) the analysis of the translation strategies as regards the passive form in EU legislation translated from English into Italian provided evidence of a lower frequency of the passive in the target texts. Interestingly, the study also revealed that the use of the passive form in the corpus of translated EU legislation was also lower compared to the reference corpus of Italian national legislation.

2. Materials and method

The EURO-CoL corpus is a "multilingually comparable corpus" (Hansen-Schirra and Teich 2009, 1162) compiled by the author. It is composed of a bilingual parallel corpus of EU legislation and a comparable corpus of Italian national legislation. The bilingual parallel corpus comprises a subcorpus of EU legislation in English (ENGLEX – 2,937,323 words) and a subcorpus of the same legislation in Italian (ITALEX – 3,018,633 words).[7] The 205 laws contained in the ENGLEX/ITALEX subcorpora are all secondary legislation (112 regulations, 78 directives and 15 decisions) adopted in the period between 2005 and 2015 and pertain to the field of consumer law.

The comparable monolingual corpus of Italian national legislation (LEGITALIA – 2,573,468 words) comprises 245 Italian laws (230 *leggi* and 15 *decreti legge*).[8] Only legislation that is unrelated to EU law was included in the reference corpus, so as not to vitiate the analysis; since most consumer legislation in Italy is based on EU directives and regulations, the Italian legislative texts included in LEGITALIA necessarily pertain to other branches of the law as well. However, the comparability of LEGITALIA with the ITALEX subcorpus was ensured by including the same type of legislation (secondary legislation), by considering the same time frame (2005-2015) and by including the integral texts of the Italian laws.

The analysis was carried out both quantitatively and qualitatively. For the quantitative analysis, the Wordsmith Tools 7.0[9] (Scott 2016) concordancing programme was used in order to extract data concerning the number of tokens and types, standardised type/token ratio and mean sentence length in the EURO-CoL corpus. Since previous corpus-based studies have shown that the different types of law present specific characteristics and features (Caliendo 2007; Biel 2014; Seracini 2020), the data was also extracted considering the directives, regulations and decisions included in ENGLEX and ITALEX separately.

On the basis of the quantitative data, the corpus was investigated further with a qualitative approach. The research on explicitation focused on shifts between the ENGLEX and ITALEX subcorpora at a lexical and morphosyntactic level. At a lexical level, terms related to the field of consumer protection law were considered. In particular, "purely technical terms" (i.e. terms used in the legal sphere only) and "semi-technical terms" (i.e. terms found in other contexts as well, but which are used in a legal context with a different meaning) were investigated (Alcaraz and Hughes 2014, Chapter 5). At a morphosyntactic level, the analysis took into consideration the translational patterns for deontic modal verbs (cf. Seracini 2020).

The qualitative analysis of simplification focused on shifts at a morphosyntactic, stylistic and syntactic level between the two subcorpora. In particular, the analysis considered the translational patterns for the negative form, the passive voice, modality, and the theme/rheme relation.

In the case of high frequency words in the corpus (e.g. 'shall'), where the analysis of the translational patterns of all the occurrences was not possible, the corpus linguistics method of "hypothesis testing" (Hunston 2002, 52) was adapted to the purposes of research on translation. With the "hypothesis testing" method, "a small selection of lines is used as a basis for a set of hypotheses about patterns," while "[o]ther searches are then employed to test those hypotheses and form new ones" (Hunston 2002, 52). The method was adapted to the study of the translational patterns by considering a set of thirty concordance lines containing the linguistic element under investigation from ENGLEX and examining the translational pattern of each concordance line in ITALEX. This procedure was repeated several times by considering a set of thirty concordance lines at a time, until no new translational patterns emerged.

3. Results and discussion

A quantitative analysis was carried out with the aid of the Wordsmith Tools 7.0 (Scott 2016) concordancing programme in order to calculate the number of tokens, types, standardised type/token ratio and mean sentence length in the EURO-CoL corpus. Table 1 below presents the results of the analysis.

The same analysis carried out by subgenre, considering directives, regulations and decisions separately, provided the results presented in Table 2 below. As Table 2 shows, directives

Table 1: Data from the EUROCoL corpus.

	ENGLEX	ITALEX	LEGITALIA
Number of files	205	205	245
Tokens	2,937,323	3,018,633 (+2,8%)	2,573,468
Types	48,704	65,561	35,034
Type/token ratio	1,66	2,17	1,46
Standardised type/token ratio	26.63	31.17 (+17%)	33.89
Mean sentence length (words)	43.68	44.11 (+1%)	43.45

are the subgenre that presents the highest increase (+5%) of tokens in ITALEX. However, the differences between the three subgenres in terms of variation in the number of tokens are very slight. Sections 3.1 and 3.2 discuss the quantitative data with regard to explicitation and simplification respectively and report on the results of the qualitative analysis.

Table 2: Data from ENGLEX and ITALEX by type of law.

	ENGLEX			ITALEX		
	Directives	Regulations	Decisions	Directives	Regulations	Decisions
Tokens	749,560	2,129,222	58,541	785,066 (+5%)	2,174,519 (+1%)	59,048 (+0,9%)
Types	19,683	43,964	3,521	23,013	58,131	5,027
STT ratio	30.24	25.68	25.75	35.53 (+17%)	29.79 (+16%)	29.88 (+16%)
Mean SeL	38.94	45.46	40.6	38.42 (-0.5%)	46.48 (+2%)	38.95 (-4%)

3.1 Explicitation in the EUROCoL corpus

As Table 1 shows, the number of running words in the translated corpus is slightly higher in ITALEX (+2.8%) compared to ENGLEX. This slight increase could be considered, in principle, an indication of the explicitation hypothesis, i.e. the translators' tendency to make explicit what is implicit in the original texts. However, as reported in section 1.1, quantitative data is not sufficient to provide evidence of explicitation; the corpus was, therefore, investigated qualitatively, in order to shed light on the various features that are responsible for a higher number of words in the translated legislation compared to the original laws. Differences between the two language systems as possible causes for an increase in the number of words in the ITALEX subcorpus were first considered. At a lexical level, verbal nouns translated with a corresponding noun phrase in Italian were found (example 1 from Decision 2006/28/EC).

(1) Article 3 Tagging

Articolo 3 Apposizione dei marchi

The analysis also revealed that the structure of the English noun phrase (noun pre-modifier(s) + head noun) is used frequently in ENGLEX. As example 2 from Regulation 1223/2009 below shows, translating this morphosyntactic structure entails introducing words, such as prepositions, in the translated texts.

(2) Ensuring traceability of a cosmetic product throughout the whole supply chain helps to make **market surveillance** simpler and more efficient. An **efficient traceability system** facilitates **market surveillance authorities'** task of tracing economic operators.

Garantire la rintracciabilità di un prodotto cosmetico in tutta la catena di fornitura contribuisce a semplificare la **vigilanza sul mercato** e a migliorarne l'efficienza. Un **sistema efficiente di tracciabilità** agevola alle **autorità di vigilanza del mercato** il compito di rintracciare gli operatori economici.

The examples above provide evidence of obligatory explicitation (Klaudy 1998), which indicates that some of the linguistic structures that characterise the two languages differently can be partly responsible for an increase in the number of words in the target texts compared to the source texts. However, the analysis also demonstrated that other types of explicitation characterise the translated legislation.

At a lexical level, evidence of pragmatic explicitation (Klaudy 1998) was found in the use of descriptive equivalents, which complement the translated legal term with a description that provides the recipients with the knowledge they lack in the target culture (Biel 2014, 43). As Biel (2014, 43) observes, a descriptive equivalent "usually involves explicitation, i.e. making explicit in the TT what may be implicit in the ST". In the following examples (3) and (4) from Regulation 8/2008, the legal terms 'wet lease', 'dry lease', and 'wet lease-out' are translated with a borrowing from English followed by an explanation in Italian:

(3) Leasing (a) | Terminology | Terms used in this paragraph have the following meaning: | (1) | **Dry lease** — Is when the aeroplane is operated under the AOC of the lessee. | (2) | **Wet lease** — Is when the aeroplane is operated under the AOC of the lessor.

Noleggio a) | Terminologia | I termini usati nel presente paragrafo hanno il seguente significato: | 1) | **Dry lease (noleggio a scafo nudo)** — quando l'impiego del velivolo avviene in accordo alle specifiche del COA del locatario; | 2) | **Wet lease (noleggio con equipaggio)** — quando l'impiego del velivolo avviene in accordo alle specifiche del COA del locatore.

(4) **Wet lease-out** | A Community operator providing an aeroplane and complete crew to another entity, […]

Wet lease-out (cessione a noleggio con equipaggio) | L'operatore comunitario che fornisce un velivolo con equipaggio completo ad un altro soggetto […]

Interestingly, the website of the Italian civil aviation authority, ENAC (Ente Nationale per l'Aviazione Civile), uses the English terms 'dry lease' and 'wet lease' followed by the equivalent Italian terminology, respectively 'locazione' and 'noleggio' in the webpages addressed to the lay public.[10] On the contrary, documents addressed to insiders often use the English 'dry lease' and 'wet lease' to refer to the two different types of contracts with no explanation or translation,[11] which proves that the English terms are commonly used in place of the Italian equivalents among insiders.

In the EU regulation, despite the fact that the English borrowings are defined and explained at the beginning of the legislative text (see example 3 above), the Italian descriptive equivalent is subsequently used both alongside and in place of the English term (examples 5 and 6).[12]

(5) **Dry lease-in** | (i) | A Community operator **shall not dry lease-in** an aeroplane from an entity other than another Community operator, unless approved by the Authority.

Dry lease-in (presa a noleggio a scafo nudo) | i) | Un operatore comunitario **non prende a noleggio a scafo nudo** un velivolo di un soggetto che non sia un altro operatore comunitario, salvo approvazione dell'Autorità.

(6) A Community operator shall ensure that, with regard to aeroplanes that are **dry leased-in**, any differences from the requirements prescribed in Subparts K, L, and/or OPS 1.005(b), are notified to and are acceptable to the Authority.

L'operatore comunitario assicura che, per quanto riguarda i velivoli **presi a noleggio a scafo nudo**, tutte le differenze rispetto ai requisiti di cui ai capitoli K, L e/o alla norma OPS 1 005, lettera b), sono notificate all'Autorità e da questa accettate.

In the following example (7) from Regulation 8/2008, a descriptive explanation ('addestrato in materia di gestione delle risorse dell'equipaggio') of the English acronym CRM is added in the Italian target text.

(7) (iii) | Subparagraph (a)(4)(i) applies as follows. Operator proficiency check may be conducted by a Type Rating Examiner (TRE), Class Rating Examiner (CRE) or by a suitably qualified commander nominated by the operator and acceptable to the Authority, **trained in CRM concepts** and the assessment of CRM skills.

iii) | la lettera a), punto 4.i), si applica come segue. Il controllo di professionalità da parte dell'operatore può essere condotto da un esaminatore di abilitazione per tipo (TRE), un esaminatore di abilitazione per classe (CRE) oppure da un comandante adeguatamente qualificato designato dall'operatore e accettabile per l'Autorità, **addestrato in materia di gestione delle risorse dell'equipaggio (Crew Resource Management — CRM)** e nella valutazione delle capacità sotto il profilo di CRM;

Another type of addition related to explicitation was also found at a lexical level. In specialised fields where the English language is used internationally, such as business management, the English term is sometimes added as a gloss to the Italian term, despite the fact that the Italian equivalent is commonly used. As example (8) from Regulation 859/2008 shows, the English terms 'accountable manager' and 'quality manager' accompany the Italian translation of the terms.

(8) Quality system (a) | An operator shall establish one quality system and designate one **quality manager** to monitor compliance with, and adequacy of, procedures required to ensure safe operational practices and airworthy aeroplanes. Compliance monitoring must include a feed-back system to the **accountable manager** (see also OPS 1.175 (h)) to ensure corrective action as necessary.

Sistema di qualità a) | L'operatore stabilisce un unico sistema di qualità e designa un unico **responsabile della qualità (Quality Manager)** al fine di controllare l'adeguatezza e il rispetto delle procedure richieste per garantire il sicuro svolgimento delle operazioni e l'aeronavigabilità dei velivoli. Il controllo del rispetto delle procedure deve anche prevedere un sistema per riferire le risultanze al **dirigente responsabile (Accountable Manager)** [cfr. anche la norma OPS 1.175(h)] in modo da garantire, in funzione delle necessità, l'adozione delle misure correttive.

The analysis also provided evidence of optional explicitation (Klaudy 1998), where the changes introduced in the target texts improved the naturalness of the legislative texts in Italian. In example (9) from Regulation 813/2013, the verbal structure 'when reviewing' is translated with the nominal structure 'all'atto del riesame' which is more idiomatic in Italian legal language, as data from the reference corpus LEGITALIA confirms (7 occurrences of the verb 'riesaminare' versus 76 occurrences of the noun 'riesame').

(9) The appropriateness of setting ecodesign requirements for these greenhouse gas emissions will be reassessed **when reviewing** this Regulation.

L'opportunità di stabilire specifiche per la progettazione ecocompatibile connesse a tali emissioni di gas a effetto serra sarà valutata nuovamente **all'atto del riesame** del presente regolamento.

Evidence of translation-inherent explicitation was also found in the ITALEX subcorpus both at a lexical and morphosyntactic level. At a lexical level, further specification is sometimes provided to accompany a legal term, as in example (10) below from Regulation 909/2014, where 'elevato' [high] is added in the translated text.

> (10) (d) it shall mitigate the corresponding liquidity risks with qualifying liquid resources in each currency such as cash at the central bank of issue and at other **creditworthy** financial institutions
>
> (d) attenua il corrispondente rischio di liquidità con risorse liquide di alta qualità in ciascuna valuta, come contante presso la banca centrale di emissione o altri enti finanziari **con merito di credito elevato**

Similarly, in the following example (11) from Regulation 859/2008, where the source text speaks of 'requirements applicable to synthetic training devices', the Italian translation adds 'norma' [norm] and 'JAR' (an acronym that refers to the term 'Joint Aviation Requirements' and is defined in an earlier section of the law): an addition which provides more accuracy and clarity to the Italian version of the law.

> (11) All synthetic training devices (STD), such as flight simulators or flight training devices (FTD), replacing an aeroplane for training and/or checking purposes are to be qualified in accordance with the requirements applicable to **synthetic training devices**. An operator intending to use such STD must obtain approval from the Authority.
>
> Tutti i dispositivi di addestramento (STD), quali i simulatori di volo o i dispositivi di addestramento al volo (FTD), che sostituiscono un velivolo a fini di addestramento e/o controllo devono essere qualificati conformemente ai requisiti applicabili della **norma JAR- STD**. L'operatore che intende utilizzare tali dispositivi deve ottenere l'approvazione dell'autorità.

In the following example from Regulation 8/2008 (12), the translator provides further clarification by adding 'prima, durante e dopo il volo' [before, during and after the flight] in the Italian version of the law where the English version more generically uses 'for all phases of operation of the aeroplane'.

> (12) An operator shall establish a check-list system to be used by crew members for all phases of operation of the aeroplane under normal, abnormal and emergency conditions as applicable, to ensure that the operating procedures in the Operations Manual are followed.
>
> L'operatore stabilisce un sistema di liste di controllo (check-list) che devono essere utilizzate dai membri d'equipaggio nelle varie fasi del volo **(prima, durante e**

> **dopo il volo)** in condizioni normali, non normali e di emergenza, al fine di garantire che siano osservate le procedure operative riportate nel manuale delle operazioni.

In all three cases above, the translated text explicitates what is left implicit in the English version by addition. Other translation-inherent explicitation involves specification, as can be seen in the following example (13) from Regulation 8/2008, where 'the flight' is translated in Italian as 'lo svolgimento del volo' [the course of the flight].

> (13) […] if he/she knows or suspects that he/she is suffering from fatigue, or feels unfit to the extent that the flight **may** be endangered.
>
> […] se ha la sensazione di una non perfetta efficienza fisica al punto da **poter determinare** una situazione di pericolo per lo svolgimento del volo.

At a morphosyntactic level, the analysis of the translational patterns for deontic modal verbs revealed that the translated text sometimes conveys a more restrictive meaning, providing an interpretation of the degree of obligation or permission that prevents the risk of misinterpretations. In Seracini (2020) it was found that the modal 'should' often tends to be translated with a stronger connotation of obligation, in particular – but not exclusively - when it occurs in the enacting terms of the legislation. In example (14) from Directive 2006/42/EC, the mild degree of obligation expressed by the two instances of 'should' is translated, respectively, with the present indicative of the verb 'dovere', which expresses strong obligation, and the final clause 'da rispettare', which indicates that something 'is to be done'.

> (14) the description of the adjustment and maintenance operations that **should** be carried out by the user and the preventive maintenance measures that **should** be observed;
>
> la descrizione delle operazioni di regolazione e manutenzione che **devono** essere effettuate dall'utilizzatore nonché le misure di manutenzione preventiva **da rispettare;**

Evidence of explicitation by means of specification was also found in the above-mentioned study (Seracini 2020) in the translational patterns for the modal 'shall'. An example of this can be seen in the extract below from Regulation 10/2011 (example 15), where the verb 'dovere' is used.

> (15) In a plastic multi-layer material or article, the composition of each plastic layer **shall comply** with this Regulation.
>
> La composizione di ogni strato di materia plastica di un materiale o oggetto di materia plastica multistrato **deve essere conforme** al presente regolamento.

It is significant that this choice departs from the norm concerning national legislation that recommends the use of the present indicative to express a prescriptive meaning, as specified in

the following extract from the Italian *Guida alla Redazione dei Testi Normativi*,[13] the official guidelines for the drafting of legislative texts:

> "The appropriate verb form to express obligation in the law is the present indicative. The subjunctive and the future tense do not achieve the same effect, in that the order they express is hypothetical and differed. In any case, use of modes or tenses different from the present indicative produces a text that is not homogeneous and is, therefore, avoided." [my translation]

Interestingly, in other parts of the same EU regulation, the modal verb 'shall' is translated with the present indicative, which is consistent with the above-mentioned guidelines (example 16).

(16) An additive **shall be removed** from the provisional list

Un additivo è **soppresso** dall'elenco provvisorio

A more in-depth analysis was carried out by applying the "hypothesis testing" method (Hunston 2002) on a total of 180 parallel sections where 'shall' and its translations occurred in ENGLEX and ITALEX (Seracini 2020). The analysis pointed to a frequent use of the verb 'dovere' to translate the modal 'shall' where, if a present indicative had been used, the target text could have been misinterpreted as conveying a factual or informative meaning, rather than a prescriptive meaning. Specification appears, therefore, to be aimed at reducing potential ambiguity in the target text. Example (17) from Directive 2006/141/EC illustrates this point.

(17) The labelling of infant formulae and follow-on formulae **shall be designed** to provide the necessary information about the appropriate use of the products so as not to discourage breast feeding.

Le etichette degli alimenti per lattanti e degli alimenti di proseguimento **devono essere concepite** in modo da fornire le informazioni necessarie all'uso appropriato di questi prodotti e non scoraggiare l'allattamento al seno.

In example (18) from Decision 2006/1005/EC, the two instances of the modal 'shall' are translated, respectively, with the present indicative and the future of the verb 'dovere', which makes the meaning more explicit.

(18) LCD refresh rate **shall be set** to 60 Hz, unless a different refresh rate is specifically recommended by the manufacturer, in which case that rate **shall be used**.

La frequenza di aggiornamento dei monitor a cristalli liquidi **deve essere fissata** a 60 Hz, a meno che il fabbricante non indichi espressamente una frequenza diversa, che **dovrà** in tal caso essere utilizzata.

3.2 Simplification in the EURO-CoL corpus

As Table 1 shows, there is a higher number of types in ITALEX (+33%) compared to ENGLEX. Since ENGLEX and ITALEX have a comparable number of tokens, the number of types in the two subcorpora, is, at least to a certain degree, comparable – something which would not have been the case had there been a large discrepancy in the size of the two subcorpora (cf. Kenny 2001). The higher number of types could be read as an indication of increased lexical variety in the translated laws, which would contradict the simplification hypothesis. However, it is necessary to consider a number of factors before drawing such a conclusion.

The first element to consider is that, as previously demonstrated, the Italian translated laws contain a number of semi-specialised and specialised terms in English that are added as a gloss to the Italian terms. Moreover, some technical terms are left untranslated in English, as can be seen in example (19) from Regulation 790/2009, where the international chemical identification is not translated into Italian.

> (19) 006-007-00-5 | salts of hydrogen cyanide with the exception of complex cyanides such as ferrocyanides, ferricyanides and mercuric oxycyanide and those specified elsewhere in this Annex
>
> 006-007-00-5 | salts of hydrogen cyanide with the exception of complex cyanides such as ferrocyanides, ferricyanides and mercuric oxycyanide and those specified elsewhere in this Annex

The second element to consider is that Italian is a more inflected language than English. For example, the English definite article 'the' is translated with seven distinct words in Italian (i.e. the definite articles 'il', 'lo', 'la', 'i', 'le', 'gli', 'l'), depending on the gender and number of the noun it goes with, as well as on the letter the noun starts with. All these features can be responsible for a higher number of distinct words in the Italian versions of the laws compared to the source texts. The results of the quantitative analysis were not, therefore, considered to be indicative of increased lexical variety in the translated laws.

The data also shows a 17% increase in the standardised type/token ratio in ITALEX compared to ENGLEX. This could also be affected by the higher number of types in ITALEX.[14] The three subgenres present a very similar variation, with a slightly higher percentage of increase in directives (+17%) compared to both regulations and decisions (+16%). The comparison between ITALEX and LEGITALIA is more revealing: the data shows that the standardised type/token ratio in the reference corpus of national Italian legislation is 9% higher than in the ITALEX subcorpus. This points to reduced lexical density in the translated legislation compared to the non-translated legislation, which would support the simplification hypothesis.

The data concerning average sentence length in ENGLEX, ITALEX and the LEGITALIA reference corpus was also compared. The results show a similar mean sentence length between ENGLEX and ITALEX, with a mere 1% increase in the translated legislation. The difference

between directives, regulations and decisions is negligible. The similarity between ENGLEX and ITALEX as regards mean sentence length suggests that translators tend to comply with the requirement of the so-called "sentence rule," whereby "language versions should have the same 'sentence boundaries"' [15] and are easily aligned for multilingual display.

Interestingly, the results are also similar to LEGITALIA. However, due to the differences in layout and structure between EU legislation and national Italian legislation, a direct comparison of mean sentence length between ITALEX and LEGITALIA cannot be made. Most of the EU legislative texts in the corpus, for example, contain a section that is not present in Italian legislation with the definitions of all the terms used in the law. These definitions constitute one long sentence and consequently affect the measure of mean sentence length, as the example (20) below from *Regulation 509/2006* illustrates.

(20)

Article 2

Definitions

1.For the purposes of this Regulation:

(a)| 'specific character' means the characteristic or set of characteristics which distinguishes an agricultural product or a foodstuff clearly from other similar products or foodstuffs of the same category;

(b) | 'traditional' means proven usage on the Community market for a time period showing transmission between generations; this time period should be the one generally ascribed to one human generation, at least 25 years;

(c) | 'traditional speciality guaranteed' means a traditional agricultural product or foodstuff recognised by the Community for its specific character through its registration under this Regulation;

(d) | 'group' means any association, irrespective of its legal form or composition, of producers or processors working with the same agricultural product or foodstuff.

Articolo 2

Definizioni

1.Ai fini del presente regolamento si intende per:

a)| «specificità», l'elemento o l'insieme di elementi che distinguono nettamente un prodotto agricolo o alimentare da altri prodotti o alimenti analoghi appartenenti alla stessa categoria;

b) | «tradizionale», un uso sul mercato comunitario attestato da un periodo di tempo che denoti un passaggio generazionale; questo periodo di tempo dovrebbe essere quello generalmente attribuito ad una generazione umana, cioè almeno 25 anni;

> c) | «specialità tradizionale garantita», prodotto agricolo o alimentare tradizionale la cui specificità è riconosciuta dalla Comunità attraverso la registrazione in conformità del presente regolamento;
>
> d) | «associazione», qualsiasi associazione, a prescindere dalla sua forma giuridica o dalla sua composizione, di produttori o di trasformatori che trattano il medesimo prodotto agricolo o alimentare.

Due to the difference in structure and layout, the similar measure of mean sentence length in ITALEX and LEGITALIA was not considered evidence of a tendency to conform to target culture drafting conventions on the translators' part.

The qualitative analysis carried out in Seracini (2020) provided evidence of recurrent shifts at a morphosyntactic, stylistic and syntactic level. At a morphosyntactic level, the analysis showed a tendency to omit words that do not carry a full meaning in the source text. In example (21) from Regulation 10/2011, the first occurrence of the modal verb 'should' is translated with the present indicative of the lexical verb, whereas the three subsequent instances of 'should' are translated with the Italian *verbo servile* 'dovere' which, as mentioned previously, expresses strong obligation. The translational choices reflect the function of the modal verb in the original text. In the first instance, 'should' merely has a tentative connotation, which is different from the other three occurrences, where the modal expresses a deontic meaning. The example also provides further confirmation that, as previously mentioned, 'should' is often translated with expressions that convey a stronger connotation of obligation, where this is implied in the source text.

> (21) The specific migration limit is a maximum permitted amount of a substance in food. This limit **should ensure** that the food contact material does not pose a risk to health. It **should be ensured** by the manufacturer that materials and articles not yet in contact with food will respect these limits when brought into contact with food under the worst foreseeable contact conditions. Therefore compliance of materials and articles not yet in contact with food **should be assessed** and the rules for this testing **should be set out**.
>
> Il limite di migrazione specifica corrisponde alla quantità massima di una sostanza consentita nei prodotti alimentari. Detto limite **garantisce** che il materiale destinato a venire in contatto con i prodotti alimentari non presenti rischi per la salute. Il fabbricante **deve garantire** che i materiali e gli oggetti che non sono ancora in contatto con prodotti alimentari rispetteranno tali limiti nel momento in cui entreranno in contatto con i prodotti alimentari nelle peggiori condizioni di contatto prevedibili. Di conseguenza, **deve essere valutata** la conformità dei materiali e degli oggetti che non sono ancora in contatto con i prodotti alimentari, ed **è necessario** stabilire le norme per la realizzazione di tali prove.

At a stylistic level, the above-mentioned analysis revealed a frequent tendency to reduce repetition in the translated texts (Seracini 2020). One example is provided below (example 22 from Regulation 216/2008), where the repetition of the modal 'may' is avoided in the Italian translation by means of an ellipsis. Moreover, the adjective 'own', which reinforces the idea of ownership in the source text, is omitted in the Italian translation.

(22) In each of the Member States, the Agency shall enjoy the most extensive legal capacity accorded to legal persons under their laws. It may, in particular, acquire or dispose of movable and immovable property and **may** be a party to legal proceedings. 3. The Agency may establish its **own** local offices in the Member States, subject to their consent.

L'Agenzia gode in tutti gli Stati membri della più ampia capacità giuridica riconosciuta alle persone giuridiche dalle rispettive legislazioni. In particolare può acquistare od alienare beni mobili e immobili e stare in giudizio. 3. L'Agenzia ha facoltà di istituire uffici locali negli Stati membri, se questi lo consentono

A tendency to reduce redundant clauses by means of alternative, more simplified structures also emerged from the analysis (Seracini 2020). In example (23) below from Directive 2006/42/EC the clause 'whatever they may be' is translated with the noun phrase 'qualsiasi tipo' [any type].

(23) automatic or manual stopping of the moving parts, **whatever they may be**, must be unimpeded,

l'arresto manuale o automatico degli elementi mobili **di qualsiasi tipo** non deve essere impedito,

The analysis of the parallel corpus also revealed that the syntactic structure is frequently simplified (Seracini 2020). One frequent shift introduces changes in the theme/rheme relation, which simplifies the sentence structure in the target text and makes the prescriptive principles clearer, as example (24) from Regulation 10/2011 illustrates.

(24) **Recently additional monomers, other starting substances and additives have received a favourable scientific evaluation by the Authority** and should now be added to the Union list.

L'Autorità ha recentemente effettuato una valutazione scientifica positiva di ulteriori monomeri, altre sostanze di partenza e additivi, che sarebbe quindi opportuno aggiungere all'elenco dell'Unione.

Simplification was also found in the frequent avoidance of complex negative structures in the target texts. In example (25) from Regulation 1899/2006, the structure 'may not'/'unless' is translated with the positive form 'può'/'solo se' [may/only if].

(25) A pilot **may not** continue an approach below MDA/MDH **unless** at least one of the following visual references for the intended runway is distinctly visible and identifiable to the pilot.

Il pilota **può** continuare un avvicinamento al di sotto della MDA/MDH **solo se** almeno uno dei seguenti riferimenti visivi per la pista ove intende effettuare l'atterraggio sia chiaramente visibile ed identificabile dal pilota.

Syntactic discontinuity found in the source text is also frequently avoided in the target text, which results in a more simplified sentence structure. One example is provided by the extract below (example 26) from Regulation 66/2010, where the separation between the adverb 'where' and the noun phrase 'any competent body' found in the source text is avoided in the target text ('qualora un organismo competente' [where a competent body]).

(26) **Where**, **giving the user of the EU Ecolabel the opportunity to submit observations**, **any competent body which finds** that a product bearing the EU Ecolabel does not comply with the relevant product group criteria or that the EU Ecolabel is not used in accordance with Article 9, it shall either prohibit the use of the EU Ecolabel on that product, or, in the event that the EU Ecolabel has been awarded by another competent body, it shall inform that competent body.

Qualora un organismo competente rilevi che un prodotto che reca il marchio Ecolabel UE non rispetta i criteri stabiliti per il rispettivo gruppo di prodotti o che il marchio Ecolabel UE non viene usato conformemente a quanto previsto dall'articolo 9, **dopo aver consentito all'utilizzatore del marchio Ecolabel UE di inviare le proprie osservazioni** l'organismo vieta l'uso del marchio su tale prodotto o, qualora il marchio Ecolabel UE sia stato assegnato da un altro organismo competente, informa quest'ultimo.

A simplifying tendency was also observed in the cases where passive forms are translated with active forms (Seracini 2019; 2020). As example (27) from Directive 2014/28/EU shows, the passive 'has been carried out' is transformed into the active form in the target text, and the agent ('the manufacturer') becomes the subject of the clause ('il fabbricante').

(27) Before placing an explosive on the market importers shall ensure that the appropriate conformity assessment procedure referred to in Article 20 has been carried out by the manufacturer.

Prima di immettere un esplosivo sul mercato, gli importatori assicurano che **il fabbricante abbia eseguito** l'appropriata procedura di **valutazione della conformità di cui all' articolo 2**.

4. Conclusion

The present study set out to investigate explicitation and simplification in translated EU legislation. As regards explicitation, the slight increase in the number of tokens in the subcorpus of Italian legislation provided an indication that explicititation could potentially characterise the translated legislation. The subsequent qualitative analysis confirmed that the translated texts tend to be more explicit at a lexical and morphosyntactic level. Beside the instances of obligatory explicitation, the study revealed that pragmatic explicitation is frequently found in the translation of technical and semi-technical terms where the terms in the target texts are provided both in English and in Italian by means of glosses and descriptive equivalents. It can be hypothesised that this increase in the level of explicitness in the translated texts is intended to avoid the risk of misinterpretations. Similarly, in the cases of translation-inherent explicitation identified in the corpus, there is a tendency to add linguistic items that contribute to making the translated texts more accurate. As regards optional explicitation, a number of changes that improve the naturalness of the target texts were observed, such as the use of nominal structures in place of verbal structures.

Evidence of explicitation by means of specification was also found. In the case of deontic modality, the analysis revealed that the linguistic choices in the target text sometimes make the degree of obligation stronger (e.g. 'should'), if this is implied in the original text. This results in target texts that are less likely to be misinterpreted than the source texts, even if this sometimes means going against target language conventions.

As regards simplification, the quantitative data concerning standardised type/token ratio pointed to evidence in support of the simplification hypothesis. The qualitative analysis indicated that a simplifying tendency occurs at a morphosyntactic, stylistic and syntactic level. A tendency to reduce repetition and to omit unnecessary or redundant linguistic elements in the target texts was observed. In particular, the translated texts often have more simplified sentence structures, avoid complex negative forms and use the active voice in place of the passive. As a result, the translated legislation is in some cases clearer and more readable.

Considering the emphasis of EU guidelines on clarity, readability and quality in legislation, all the results in the present study could be read merely as evidence of the fact that translators comply with EU institutional norms. For example, both the English and the Italian versions of the manual for drafters and translators, *How to Write Clearly* (European Commission 2011), [16] specify that, when drafting a document, it is preferable to use the affirmative instead of the negative form, the active instead of the passive form, and that the actions should be placed "in the order in which they occur" (European Commission 2011, 7). However, since the same guidelines apply to legal drafting, not only translation, it can be hypothesised that the instances where translated legislation is clearer, more readable and less ambiguous than the source texts also provide evidence in favour of the explicitation and simplification hypotheses in legal translation.

Notes

[1] Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1485097054576\&uri=CELEX:31958R0001(02). Accessed 30 November 2020.

[2] Information retrieved from Translation and Multilingualism. http://bookshop.europa.eu/en/translation-and-multilingualism-pbHC0414307/. Accessed 30 November 2020.

[3] Available at https://op.europa.eu/en/publication-detail/-/publication/3879747d-7a3c-411b-a3a0-55c14e2ba732. Accessed 5 December 2020.

[4] Available at http://ec.europa.eu/translation/maltese/guidelines/documents/dgt_ translation_quality_guidelines_en.pdf. Accessed 18 December 2020.

[5] Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX$\%$3A52015DC0215. Accessed 18 December 2020.

[6] "La vaghezza avrebbe il pregio della flessibilità, anziché il difetto dell'imprecisione" (Antelmi 2008, 94).

[7] The EU laws were downloaded from Eur-Lex (https://eur-lex.europa.eu/homepage.html?locale=en), the section of the EU website containing all EU legislation.

[8] The Italian laws were downloaded from *Parlamento.it* (www.parlamento.it), the official website of the Italian Parliament and from *Normattiva* (www.normattiva.it), the official website of the Italian legislation in force.

[9] Available at http://www.lexically.net/wordsmith/.

[10] Available at https://www.enac.gov.it/trasporto-aereo/compagnie-aeree/licenze-di-esercizio/licenza-di-trasporto-aereo/impiego-aeromobili. Accessed 2 December 2020.

[11] See, for example, document available at https://www.enac.gov.it/sites/default/files /allegati/ 2018-Set/77531_PROT_18_07_2014_Approvazione_impiego_aeromobili.pdf. Accessed 2 December 2020.

[12] The use of English borrowings alongside the Italian term has been observed above all in regulations. This may be explained with reference to the fact that these types of laws are applied directly within the various national legislative system. The study on Lawmaking in the EU Multilingual Environment published in 2010 by the Directorate General for Translation of the European Commission (available at https://op.europa.eu/it/publication-detail/-/publication/7db404b5-48e5-4c2b-aabd-82db6a034eab; accessed 15 December 2020) reports that "[i]n the case of regulations, the impact of national legal or technical terms on the vocabulary of the regulation is more significant than with directives because the national legislator does not have the possibility of remedying the incorrect terminology in the phase of transposition."

[13] *Gazzetta Ufficiale* n. 101 of 3. May 2001 – *Supplemento Ordinario* n. 105. Available at http://www.gazzettaufficiale.it/atto/stampa/serie$_g$enerale/originario. Accessed 3 December 2020.

[14] Kenny (2001) points out that, when data referring to standardised type/token ratio is ana-

lysed, it is necessary to take certain issues into account. Two of these issues in particular need to be considered in the present study. The first issue is that homographs are not automatically considered by the concordancer as different words. For example, the word 'list' can be both a verb or a noun. If a corpus is not annotated, the concordancer considers the verb and the noun as the same word. The second issue is that the different word forms of a lemma are counted by the concordancer as different types. This issue affects in particular the data from corpora of highly inflected languages, such as Italian. On the basis of these two issues, the data concerning standardised type/token is considered here as merely providing an indication of a general trend.

[15] Information retrieved from the *DGT Translation Quality Guidelines* document (available at http://ec.europa.eu/translation/maltese/guidelines/documents/dgt$_t$ranslation$_q$uality$_g$uidelines$_e$n.pdf; accessed 2 December 2020).

[16] Available at http://bookshop.europa.eu/en/how-to-write-clearly-pbHC3010536/. Accessed 5 December 2020.

References

Alcaraz, Enrique, and Brian Hughes. 2014. *Legal Translation Explained*. London and New York: Routledge.

Anselmi, Simona, and Francesca Seracini. 2015. "The Transposition of EU Directives into British Legislation as Intralingual Translation: A Corpus-Based Analysis of the Rewriting Process." *Textus* 28 (2): 39–62.

Antelmi, Donella. 2008. "Vaghezza, Definizioni e Ideologia Nel Linguaggio Giuridico." In *Il Linguaggio Giuridico*, ed. by Giuliana Garzone, and Francesca Santulli, 89–119. Milano: Giuffrè.

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies - Implications and Applications." In *Text and Technology*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233–250. Amsterdam and Philadelphia: John Benjamins.

Baker, Mona. 1996. "Corpus-Based Translation Studies: The Challenges That Lie Ahead." In *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, ed. by Harold L. Somers, 175–186. Amsterdam and Philadelphia: John Benjamins.

Becher, Viktor. 2010a. "Abandoning the Notion of 'Translation-Inherent' Explicitation: Against a Dogma of Translation Studies." *Across Languages and Cultures* 11 (1): 1–28.

Becher, Viktor. 2010b. "Towards a More Rigorous Treatment of the Explicitation Hypothesis in Translation Studies." *Trans-Kom* 3 (1): 1–25.

Biel, Łucja. 2010. "Corpus-Based Studies of Legal Language for Translation Purposes:Methodological and Practical Potential." In *Reconceptualizing LSP*, ed. by Carmen Heine, and Jan Engberg, 1–15. Aarhus : Aarhus School of Business, Aarhus University.

Biel, Łucja. 2014. "The Textual Fit of Translated EU Law: A Corpus-Based Study of Deontic

Modality." *The Translator* 20 (2): 1–23.

Blum-Kulka, Shoshana. 1986. "Shifts of Cohesion and Coherence in Translation." In *Interlingual and Intercultural Communication*, ed. by Juliane House, and Shoshana Blum-Kulka, 17–36. Tübingen: Gunter Narr.

Blum-Kulka, Shoshana, and Eddie A. Levenston. 1983. "Universals of Lexical Simplification." In *Strategies in Interlanguage Communication*, ed. by Claus Faerch, and Gabriele Casper, 119–139. London and New York: Longman.

Caliendo, Giuditta. 2007. "Modality and Communicative Interaction in EU Law." In *Intercultural Aspects of Specialized Communication*, ed. by Christopher N. Candlin, and Maurizio Gotti, 241–259. Bern: Peter Lang.

Chesterman, Andrew. 2004a. "Beyond the Particular." In *Translation Universals*. Do They Exist?, ed. by Anna Mauranen, and Pekka Kujamäki, 33–49. Amsterdam and Philadelphia: John Benjamins.

Chesterman, Andrew. 2004b. "Hypotheses about Translation Universals." In *Claims, Changes and Challenges in Translation Studies*, ed. by Gyde Hansen, Kirsten Malmkjær, and Daniel Gile, 1–14. Amsterdam and Philadelphia: John Benjamins.

Chesterman, Andrew. 2010. "Why Study Translation Universals?" *Acta Translatologica Helsingiensia* 1: 38–48.

Chesterman, Andrew, and Rosemary Arrojo. 2000. "Shared Grounds in Translation Studies." *Target* 12 (1): 151–160.

Engberg, Jan, and Dorothee Heller. 2008. "Vagueness and Indeterminacy in Law." In *Legal Discourse across Cultures and Systems*, ed. by Vijay K. Bathia, Christopher N. Candlin, and Jan Engberg, 145–168. Hong Kong: Hong Kong University Press.

European Commission. 2011. *How to Write Clearly*. Luxembourg: Publications Office of the European Union. http://bookshop.europa.eu/en/how-to-write-clearly-pbHC3010536/.

European Commission Directorate-General for Translation. 2015. *DGT Translation Quality Guidelines*. http://ec.europa.eu/translation/maltese/guidelines/documents/dgt$_{t}$ranslatio n$_{q}$uality$_{g}$uidelines$_{e}$n.pdf.

European Union. 2015. *Joint Practical Guide of the European Parliament, the Council and the Commission for Persons Involved in the Drafting of European Union Legislation*. Luxembourg: Publications Office of the European Union. http://eur-lex.europa.eu/content/techleg/KB0213228ENN.pdf.

Faber, Dorrit, and Mette Hjort-Pedersen. 2013. "Expectancy and Professional Norms in Legal Translation: A Study of Explicitation and Implicitation Preferences." *Fachsprache* 35 (1–2): 42–62.

Hansen-Schirra, Silvia, and Elke Teich. 2009. "Corpora in Human Translation." In *Corpus Linguistics. An International Handbook. Vol. 2*, ed. by Anke Lüdeling, and Merja Kytö, 1159–1175. Berlin: de Gruyter.

Hjort-Pedersen, Mette, and Dorrit Faber. 2010. "Explicitation and Implicitation in Legal Translation - A Process Study of Trainee Translators." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 55 (2): 237–250.

House, Juliane. 2008. "Beyond Intervention: Universals in Translation." *Trans-Kom* 1 (1): 6–19.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation. A Corpus-Based Study*. Manchester: St Jerome.

Klaudy, Kinga. 1998. "Explicitation." In *Routledge Encyclopedia of Translation Studies*, ed. by Mona Baker, and Gabriela Saldanha, 2nd ed., 80–84. London and New York: Routledge.

Krogsgaard Vesterager, Anja. 2017. "Explicitation in Legal Translation — a Study of Spanish-into-Danish Translation of Judgments." *The Journal of Specialised Translation* 27: 104–123.

Laviosa, Sara. 2003. "Corpus and Simplification in Translation." In *Translation Translation*, ed. by Susan Petrilli, 153–162. Amsterdam and New York: Rodopi.

Mauranen, Anna. 2007. "Universal Tendencies in Translation." In *Incorporating Corpora. The Linguist and the Translator*, ed. by Gunilla M. Anderman, and Margaret Rogers, 32–48. Clevedon: Multilingual Matters.

Munday, Jeremy. 1998. "A Computer-Assisted Approach to the Analysis of Translation Shifts." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 43 (4): 542–556.

Olohan, Maeve, and Mona Baker. 2000. "Reporting 'that' in Translated English. Evidence for Subconscious Processes of Explicitation." *Across Languages and Cultures* 1 (2): 141–158.

Pontrandolfo, Gianluca. 2019. "Corpus Methods in Legal Translation Studies." In *Research Methods in Legal Translation and Interpreting: Crossing Methodological Boundaries*, ed. by Łucja Biel, Jan Engberg, Rosario Martín Ruano, and Vilelmini Sosoni, 13–28. London and New York: Routledgè

Pontrandolfo, Gianluca. 2020. "Testing out Translation Universals in Legal Translation: Quantitative Insights From a Parallel Corpus of Spanish Constitutional Court's Judgments Translated into English." *Comparative Legilinguistics - International Journal for Legal Communication* 43: 17–55. https://doi.org/http://dx.doi.org/10.14746/cl.2020.43.2.

Prieto Ramos, Fernando. 2014. "Legal Translation Studies as Interdiscipline: Scope and Evolution." *Meta: Journal des Traducteurs/Meta: Translators' Journal* 59 (2): 260-277.

Scott, Mike. 2016. WordSmith Tools version 7. Stroud: Lexical Analysis Software.

Séguinot, Candace. 1988. "Pragmatics and the Explicitation Hypothesis." *TTR: Traduction, Terminologie, Rédactiń*1 (2): 106–113.

Seracini, Francesca. 2019. "Simplifying EU Legislative Texts: The Contribution of Translati-

on." In *World of Words: Complexity, Creativity, and Conventionality in English Language, Literature and Culture, Vol. 1*, ed. by Veronica Bonsignori, Gloria Cappelli, and Elisa Mattiello, 325–336. Pisa: Pisa University Press.

Seracini, Francesca L. 2020. *The Translation of European Union Legislation: A Corpus-Based Study of Norms and Modality*. Milano: LED.

Toury, Gideon. 1979. "Interlanguage and Its Manifestations in Translation." *Meta : Journal des Traducteurs/Meta: Translators' Journal* 24 (2): 223–231.

Tymoczko, Maria. 1998. "Computerized Corpora and the Future of Translation Studies." *Meta: Journal Des Traducteurs* 43 (4): 652-660. https://doi.org/10.7202/004515ar.

Ulrych, Margherita. 2014. *Traces of Mediation in Rewriting and Translation*. Milano: EDUCatt.

Vinay, Jean-Paul, and Jean Darbelnet. [1958]1995. *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam and Philadelphia: John Benjamins.

English for a Global Readership: Implications for the L2 Translation Classroom

Dominic Stewart

Address: Department of Humanities, University of Trento, Italy
E-mail: dominic.stewart@unitn.it
Correspondence: Dominic Stewart
Citation: Stewart, Dominic. 2021. "English for a Global Readership: Implications for the L2 Translation Classroom." *Translation Quarterly* 101: 93-112.

Abstract

When English is the target language and the intended readership is international, the question of what is appropriate or inappropriate as a translation solution can be complex. Usage which has traditionally been considered unacceptable may come to be regarded in a different light owing to the fact that on a global level English is a moving and constantly evolving target. This state of flux represents a conundrum for translation trainers and trainees alike, whose task is however facilitated by careful use of dictionaries and language corpora. With the help of corpus data, this paper examines and discusses examples of solutions – some contentious, some more straightforward – drawn from the Italian-to-English translation classroom.

1. Background

In Stewart (2013, 225-229) I considered whether the following translation into English of a sentence in an Italian tourist brochure could be considered admissible in an L2 translation classroom if the envisaged readership is international:

[…] *a panoramic road which allows to enjoy fantastic sceneries*

This is a close and rather wooden rendering of the Italian […] *una strada panoramica che consente di godere di fantastici scenari* (literally 'a road panoramic which allows to enjoy of fantastic sceneries'). At first glance the translation might seem innocuous enough, but in reality it contains at least three potential flashpoints for classroom discussion:

> -firstly, *scenery* has traditionally been regarded as an uncountable noun, whatever the meaning. It is still listed as uncountable in British-based learner's dictionaries, and there is only one occurrence of the plural in the *British National Corpus*, whereas there are 755 occurrences of the singular, almost all uncountable (see section 3.1.3 for further details)
>
> -secondly, *panoramic road* is an unusual combination, barely attested in the *British National Corpus* and in dictionaries
>
> -thirdly, '*ALLOW to enjoy*' does not respect the local grammar associated with the active use of *allow* as outlined in teaching materials, which advocate '*ALLOW* NOUN/PRONOUN to VERB' when the verb is active, e.g., *which allows you to enjoy, which allows visitors to enjoy*.

One wonders, however, how fruitful such classroom discussions really are. The text in question was intended for an international readership, and consequently for both native and non-native speakers of English. The great majority of international readers of texts for tourism are non-native speakers of English, and one imagines that a very modest percentage of them would notice anything unidiomatic about the above sentence. Further, the translation has the virtue of being accurate and unequivocal, i.e., it captures the message of the Italian source text and the meaning is clear. In addition, the tone and register of the fragment are unobjectionable. In short, the translated fragment is serviceable and presumably useful to the reader, whether a native or non-native speaker of English. So why address these 'flashpoints' in the classroom at all? The translated text would appear to carry out its purpose, therefore why not just accept it and carry on? In any case, the large uk-domain corpus *British Web 2007* contains over 60 relevant occurrences of *sceneries*, 6 relevant occurrences of *panoramic road*, and both the *British National Corpus and British Web 2007* contain an abundance of examples of active '*ALLOW* to VERB', though the majority of them are imperative uses in recipes or instruction manuals, e.g., *bring to the boil and allow to simmer until cooked*.

2. English as a Lingua Franca

The questions raised here relate to issues concerning English as a lingua franca (ELF), a vehicular language whose main focus is the clarity and coherence of the message for people of different languages and cultures. Ife (2005, 286) summarises ELF as "a language used as a common language by speakers whose mother tongue it is not", while Jenkins' (2007, 1) definition is: "a contact language used among people who do not share a first language, and is commonly understood to mean a second (or subsequent) language of its speakers". The

description offered by Taviano (2010, ix) gives greater prominence to the involvement of both native and non-native speakers: "a contact language used mainly, though not exclusively, by non-native speakers".

Despite a plethora of publications on ELF, whether in professional or pedagogical settings (see for example Crystal (2007), Gagliardi and Maley (2010), Seidlhofer (2011), Bowles and Cogo (2015), Jenkins et al. (2018), and the *Journal of English as a Lingua Franca* founded in 2012), comparatively little attention has been devoted to ELF in the translation classroom, the main contributions being Taviano (2010, 65-87) and a special issue of the journal *The Interpreter and Translator Trainer* in 2013 (issue 7 (2)). Problematic in this field of research is that ELF has misty borders, with the result that it is often hard to establish which usage lies within the purview of ELF and which usage does not, and this is perhaps why its place in trainee translation programmes is potentially contentious. According to Taviano (2013, 162-163):

> students should also be trained to translate into ELF, that is to say to produce texts tailored to an international readership, since professional translators can be and often are commissioned to do so. The heterogeneity of international audiences, which inevitably include both native and non-native speakers – just as authors of academic papers may or may not be non-native speakers of English – makes the role of those writing or translating for such audiences particularly challenging.

This scenario is indeed 'particularly challenging', in part because translating into ELF and producing texts tailored to an international readership do not necessarily dovetail (trainees may produce a good translation into English for an international readership without ever having heard of ELF), but above all because even if trainees do aspire to adopt ELF as their target, is it indicated anywhere in dictionaries, grammars, textbooks etc. that active '*ALLOW to* VERB' is likely to be an appropriate or inappropriate sequence in ELF? As stressed by Maley (2010, 36), teachers of English as a second language tend to refer to established varieties of English because

> there are no substantive models or materials available to them, were they to wish to change in the direction of ELF. Even were they broadly supportive of the ELF concept, what precisely would the practical implications be for their teaching, other than a vaguely-formulated, more tolerant attitude towards learner 'errors'?

For further comments see Anderman and Rogers (2005) and Buckledee (2010). It is indisputable that however willing trainers and trainees may be to embrace the requirements of a global readership, both trainer and trainee require reasonably well-marked borders within which to operate, and it might as a result seem simpler in the translation classroom to adopt as the target language a native variety of English which is internationally recognized and, more important, well-documented (Stewart 2013). This position, too, however, is problematic. As a

teacher of Italian-to-English translation to Humanities students in Italy, over the years I have broadly adopted British English in the classroom when the target is international, but within this framework I feel more uneasy than ever about penalising usage which, though readable and unambiguous, is unidiomatic to native-English ears. A further complication is that even the borders of British English are misty, in part – ironically enough – because the variety in question is so well-documented, with a whole host of grammars, dictionaries corpora, text-books etc., which of course may furnish conflicting information. The following section will provide examples of the scenarios considered here. The usage discussed will again consist of renderings by translation trainees of fragments of texts in the tourism sector destined for international readers, subdivided into issues of grammar, phraseology and colligation, with summaries at the end of each section. Reference will be made to both the *British National Corpus* and the *British Web 2007* [1].

3. Examples of trainees' translations

3.1 Grammar

3.1.1 *only if/by* + no inversion of subject and verb in the following main clause

Student translations:

- *only if you go on foot you will be able to appreciate the view*
- *only by leaving the valley floor you can reach Badia di San Bartolomeo*

Grammars of English inform us that sentences introduced by *Only if* require subject-verb inversion in the following main clause, some with auxiliary verbs, and this is certainly backed up by the *British National Corpus* phrase query 'Only if you' (with initial upper case):

Only if you want to go the whole way and produce typeset quality data ***will you ever need*** *to consider anything better than VGA*

Only if you do that ***will you be*** *able to say with confidence that I am wrong*

Only if you have done a very elaborate and expensive piece of research ***will you have gone*** *beyond this to the sort of detailed description outlined above*

Only if we teach our children anxiety ***do they begin*** *to move fearfully through life*

The query 'Only by' + verb at R1 or R2, again in sentence-initial position, retrieves 75

occurrences in the *British National Corpus*, all of which are followed by SV inversion in the following main clause, for example:

Only by improving social and economic conditions ***can good health be achieved***

Only by steadily improving efficiency ***would Britain win*** *and keep its share of the world's markets.*

Only by so doing ***can their business activity generate*** *an adequate return*

The same two searches produce similar outcomes in the *British Web 2007*, albeit with much greater frequencies. Naturally the above queries capture only sentence-initial instances (and do not include other analogous usage such as *Only then* and *Only now*), but samples in both corpora suggest that SV inversion is used with barely any exceptions when these structures are clause-initial.

3.1.2 *whole* (adj) + uncountable noun

Student translation:

this attracted a lot of attention in the whole Europe

In grammars we read that *whole* as an adjective with the sense of *entire* cannot qualify uncountable nouns (**whole traffic, *whole music, *whole Ireland*), and this too is corroborated by *British National Corpus* samples, once one has excluded irrelevant occurrences such as *the whole Scotland team*, where *whole* qualifies the countable noun *team*. The *British National Corpus* has no relevant instances of, for example, *whole Europe, whole England, whole furniture, whole information, whole equipment, whole scenery*. Although there are certainly occurrences of *whole* qualifying uncountable nouns in the corpus, in these cases the adjective usually has a different meaning (*whole wheat, whole milk*).

The *British Web 2007*, however, tells a different story: the sequence *whole Europe* is attested 14 times with 12 of them relevant, *whole equipment* shows 8 relevant cases out of 10, *whole scenery* has 4 out of 4, and *whole furniture* 7 out of 7. Less prolific are *whole information* with 5 out of 21, while *whole England* has zero relevant cases out of a total of 14. This data suggests that *whole* + uncountable noun is more widespread in the *British Web 2007* than in the *British National Corpus*, from which we can perhaps conclude that this grammatical feature is more tolerated in modern web and/or non-native English.

Worth noting is that the noun *attention*, despite being listed as uncountable in dictionaries and grammars, is qualified by *whole* 16 times in the *British National Corpus* and 60 times in

the *British Web 2007*, almost always preceded by a possessive. It would therefore appear to be an exceptional case. Here are some random occurrences in the *British National Corpus*:

*For the first time she sat down and gave him **her whole attention**.*

*You dealt with one thing at a time, gave it **your whole attention**, decided it, then put it aside.*

*Harriet Tremayne became more and more of a recluse, **her whole attention** concentrated on her granddaughter.*

3.1.3 Traditionally uncountable nouns used as countable nouns

Student translations:

- in 2005, however, a lightning destroyed the bell tower

- evidences suggest that the cave was used by humans

- the church's strategic position on a rocky slope offers a stunning scenery of the surrounding mountains

Once again, the question regards the countable / uncountable usage of nouns: *lightning, evidence* and again *scenery* in the examples above. These are listed as uncountable by dictionaries, some of which include assistance on how to indicate countability when adopting nouns such as these, for example *a lightning bolt, flashes of lightning; pieces of evidence, a shred of evidence*.

- lightning

The *British National Corpus* contains very few examples of countable *lightning* in the meteorological sense, though we find a handful of plural occurrences with figurative meanings in literary texts:

*no matter how, no matter what the **lightnings** that assailed him*

***lightnings** of pain sheathed every nerve in her body*

***an invisible lightning** leaped between them*

A similar scenario is identifiable in the *British Web 2007*: countable occurrences of *lightning* refer either – with initial upper case – to a type of aeroplane or have a metaphorical, rhetorical quality:

there is an excellent chapter on his time trials flying ***Lightnings*** *with the Air Fighting Development Squadron and as a Flight Commander on 111 Squadron*

the jungle was a vivid green blanket in which rivers made silvery forked ***lightnings***

yea he sent out his arrows & scattered them, & he shot out ***lightnings*** *and discomfitted them*

and there sprang in the gloom of his soul ***a sudden lightning*** *of joy*

There are, however, exceptional cases where the countable use does refer literally to a flash of lightning:

whilst high-power surges such as ***a lightning*** *can cause immediate damage by 'frying' circuits and melting plastic and metal…).*

- *evidence*

The case of *evidence* is rather different. Despite being classified as uncountable in the major dictionaries of English (though the *Oxford Advanced Learner's Dictionary* concedes that "in academic English the plural *evidences* is sometimes used"), it is attested fairly frequently as a countable noun in both the corpora adopted here, above all in the plural: *evidences* has 29 relevant occurrences in the *British National Corpus*, and well over 1,000 in the *British Web 2007*. The sequence *an evidence of* is also recurrent: 5 in the *British National Corpus* and 114 in the *British Web 2007*.

- *scenery*

The *British National Corpus* contains no instances of the exact sequence *a scenery*, but countable uses do turn up:

First, spilitic lavas were extruded, layer upon layer, and weathered to produce ***a staircase-like scenery*** *(or "trap topography")*

The Sierra mountain range which runs the length of the north west coast of Majorca, gives the island ***a dramatic scenery*** *in contrast to the soft beaches below*

As mentioned in section 1, over 60 instances of *sceneries* are retrieved in the *British Web 2007*, while *scenery* occurs as a singular countable noun between 70 and 80 times, for example:

this is the most wonderful course with ***a scenery*** *you only can expect in heaven*

the first series of the heart-warming Monarch of the Glen, set amongst ***a glorious Highland scenery*** *and an all-star cast*

the name of Gstaad has become synonymous with the idea of sophisticated holidays in ***an unspoiled scenery***, *spectacular in winter and summer.*

over a high route that offers ***an ever-changing fantastic scenery*** *and which spares the finale – the sight of the Matterhorn towering above Zermatt – right to the last day*

The countable/uncountable issue is a stumbling block for both translation trainers and trainees. Perhaps owing to the influence and pressure of other languages, English now appears to show greater flexibility with regard to countability, the most obvious example perhaps being the noun *research*, traditionally uncountable but now tolerated in certain contexts by dictionaries such as the *Oxford Advanced Learner's Dictionary* and the *Longman Dictionary of Contemporary English*. While corpus data suggests that nouns such as *traffic* (with reference to transport), *fun* and *furniture* cling on to their uncountable status, there are currently a number of borderline cases during what appears to be a period of transition. This situation clearly engenders uncertainty, and one imagines varying degrees of tolerance on the part of translation trainers during the evaluation process.

3.1.4 Summing up

Grammatical features such as those examined above can be problematic for trainers and trainees alike, in part because conflicting information is present in language resources. In the *British National Corpus*, the adjective *whole* almost always qualifies countable nouns – notwithstanding exceptions such as *attention* – and this ties in with indications supplied in dictionaries, but in the *British Web 2007* there is a significant number of instances where *whole* qualifies uncountable nouns. The situation is even more nebulous regarding the host of nouns considered uncountable in dictionaries being used countably: there is significant evidence of this phenomenon in both corpora, but of course there are hugely different scenarios from one noun to the next. However, even dictionaries supply contrasting indications: as mentioned above, *evidences* is sanctioned by the *Oxford Advanced Learner's Dictionary* in academic contexts, but in the *Macmillan Dictionary* and the *Longman Dictionary of Contemporary English* the user is explicitly instructed never to use *evidence* countably. The situation regarding clause-initial *only if/by* is considerably clearer in language resources, which show barely any exceptions to the SV inversion rule. Having said that, the normal SV word order does not sound particularly bad and its meaning remains perfectly comprehensible. Indeed by an irony, the avoidance of the inversion in the main clause may well make the text more intelligible to the non-native English reader, who might otherwise construe the inversion as an interrogative.

See Mossop (2006, 792) and Stewart (2013, 229-232) for discussion of scenarios such as this. Let us now move on to phraseology.

3.2 Phraseology

3.2.1 *at the feet of*

Student translation:

at the feet of the mountain lies the town of Rivoli

In the *British National Corpus* and *British Web 2007* the sequence *at the feet of* is generally followed by people in power and/or religious figures (*at the feet of many masters, at the feet of Jesus*), but not by *mountains, hills, stairs, page* etc., which lie within the compass of *at the foot of* rather than *at the feet of*. Sequences of the type *at the feet of the mountain* are common in student translations from Italian to English because the Italian equivalent has a plural noun (*ai piedi della montagna*), but the information present in English dictionaries and the two corpora does not suggest any exceptions to the rule.

3.2.2 Sequences of the type *20 kilometres far, 10 miles far*

Student translations:

- *the island is five kilometres far from Grado*
- *six km far from Portoferraio, the villa lies at the foot of Mount San Martino*

English-language teaching materials inform us that whereas it is normal to state that a road is 10 yards wide, that a swimming-pool is 50 metres long, and that a river is 5 metres deep, it sounds unidiomatic to state that a town is x km far, or x km far from somewhere. Data from the *British National Corpus* and *British Web 2007* appear to consolidate this recommendation: the *British National Corpus* has no trace of the combinations *metres far, meters far, kilometres far, kilometers far, km far, miles far* etc. except within a name (Miles Far East Company), while the *British Web 2007* shows 3 occurrences of *miles far* (though two of these are part of *far away and far and wide* respectively), 9 occurrences of *kilometres far / kilometers far*, and 19 of *km far / kms far*. Therefore, once again the *British National Corpus* occurrences are in line with indications offered in teaching materials, whereas the *British Web 2007* contains some exceptions.

3.2.3 *in the last years*

Student translation:

-in the last years there have been many local initiatives

The phrase *in the last years* occurs 73 times in the *British National Corpus*: 61 of these are followed by *of*, indicating the closing years of a specific period (*in the last years of Victoria's reign, in the last years of his life*), while 4 are followed by *before* (all 4 describe the period preceding a war, e.g., *in the last years before the war*). Of the remaining 8 occurrences, only two appear to mean *in recent years*, which is the sense required in the example above:

They're ringing in the changes at British Telecom's new residential training centre in Milton Keynes. Wimpey Construction UK, Eastern has just completed the £25 million centre, the largest building project to be completed in Milton Keynes ***in the last years.***

This is a truth that Dick Lucas and the Proclamation Trust in London have effectively and consistently brought to the church ***in the last years.***

On the other hand, the longer sequence *in the last few years* occurs 199 times in the *British National Corpus*, and of these 194 seemingly correspond to *in recent years*, while only 5 correspond to *in the final years* (e.g., *but neither programme was redirected successfully towards the cities in the last few years of the 1974-9 Labour administration*). Therefore, in the *British National Corpus* there is evidence of a reasonably neat distinction: *in the last years* almost always has the meaning *in the final/closing years*, while *in the last few years* almost always means *in recent years*. The figures in the *British Web 2007*, however, are very different. Of the 399 hits of *in the last years* in this corpus, as many as a third of these can be construed with the meaning of *in recent years*. This striking discrepancy between the two corpora would suggest either a diachronic development (the *British Web 2007* is more recent) or a more widespread use of the *in recent years* interpretation in web contexts and/or among non-native speakers. Interestingly, the avoidance of *in the last years* in the sense of *in recent years* suggested by data in the *British National Corpus* is not generally reported or even hinted at in most grammars and dictionaries. Once again, trainers and trainees need to act as information managers in order to make sense of the diverging data available to them.

3.2.4 *the world war I*

Student translations:

during the World War I; at the end of the world war II

First of all, let us consider the frequencies of the following clusters in the *British National Corpus*, with no distinction of upper/lower case. By way of example I focus on sequences

denoting the first of the two world wars:

the first world war	975
world war I (Roman numeral)	255
world war one	79
the 1st world war	15
world war 1 (Arabic numeral)	11
the world war I (Roman numeral)	5 (1 relevant)
the world war one	4 (0 relevant)
the world war 1 (Arabic numeral)	0

It is noticeable that in the *British National Corpus*, out of the total number of 345 occurrences of the sequences *world war I / world war one / world war* 1 (255+79+11) only one of these features a relevant colligation with the definite article:

But the greatest blow to the dictates of fashion on women's dress came with ***the World War I****. Although Laura Ashley had*

In the remaining cases the article colligates with a following noun, for instance:

The traditional British verve for raiding had been restored after too many years under the shadow of the ***World War I failures*** *at Gallipoli.*

We started at the ***World War I Memorial***, *clearly the object of great respect and attention.*

This is a play for two characters, Siegfried Sassoon and Wilfred Owen, the ***World War One poets***, *and it would obviously be a good thing to read some of their poetry*

provides a 'period' look at early fighting aircraft which will be useful for the ***World War One aircraft buff.***

In short, and simplifying somewhat, information from the *British National Corpus* points to the fact that *first/1st world war* requires a preceding definite article, whereas *world war one/I/1* does not. *British Web 2007* searches, on the other hand, generate the following outcomes:

the first world war	10,332
world war I (Roman numeral)	3,019
world war one	1,696
the 1st world war	151
world war 1 (Arabic numeral)	775
the world war I (Roman numeral)	106 (35 relevant)
the world war one	63 (10 relevant)
the world war 1 (Arabic numeral)	29 (7 relevant)

In the top group, the main divergence from the *British National Corpus* data is that *world war 1* (775) is a lot more frequent than *the 1st world war* (151), but my interest again focuses on the bottom group. Here we discover that out of the total number of 5490 occurrences of the sequences *world war I / world war one / world war 1* (3019+1696+775), 52 of these feature a relevant colligation with the definite article, for example:

> *This story, set in **the World War I**, captures the memories of a young soldier as he looks back over his life whilst at the front*
>
> *from suicide and homosexuality to Miss Lizzy's flirtation with the suffragette movement and the horrors of **the World War I.***
>
> *Considered by many to be not only the best combatant story of **the World War I** but the best American war book since The Red Badge of Courage.*
>
> *Dick won the Military Medal in France as a stretcher bearer in **the World War One**.*
>
> *the Headlam-Morley library of books, pamphlets and manuscripts relating to **the World War 1** and the Peace Treaty, the Henry Morris collection of Irish material*

Therefore, whereas in the *British National Corpus* relevant occurrences of the definite article with *world war I / world war one / world war 1* are very close to zero, in the *British Web 2007* they amount to almost 10%.

3.2.5 Summing up

Like *only if/by* in section 3.1.1, *at the feet of* is relatively unproblematic in that the information provided in dictionaries and the two corpora is consistent: for this reason, a sequence

of the type *at the feet of the hill* – which does not sound right at all despite its semantic transparency – is presumably to be considered inappropriate. However, as in section 3.1, if we cast our net around we discover that other usage is not quite so straightforward along the axis of acceptability / non-acceptability. Combinations of the type *x kilometres far* are absent in the *British National Corpus* but do turn up in the *British Web 2007*, albeit infrequently. The phrase *in the last years* with the meaning *in recent years* is barely present in the *British National Corpus* but is recurrent in the *British Web 2007*, and relevant use of the definite article with *world war I / world war one / world war 1* is practically non-existent in the *British National Corpus* but again fairly recurrent in the *British Web 2007*. Once again, the conundrum is how trainers and trainees should react to this type of conflicting information.

3.3 Colligation

3.3.1 *view on the sea/mountains/bay*

Student translation:
there is a spectacular view on the stunning mountains

Dictionaries furnish examples of the combination *view/s on* followed by topics and themes (*issues, marriage, education*), but not by *mountains, sea* etc. From this the user can conclude that when *view* colligates with the preposition *on* it cannot have the meaning *panorama* – in this latter sense *view of* or *view over* are required[2]. This position is backed up by *British National Corpus* data: in this corpus there are, for instance, no occurrences corresponding to the simple (lemmatized) queries 'view on the mountain', 'view on the bay', 'view on the city' and just one corresponding to 'view on the sea'. The *British Web 2007* produces fairly similar outcomes, though there is a small number of relevant instances retrieved by the simple queries 'view on the mountain' (1), 'view on the bay' (2), 'view on the city' (4) and above all 'view on the sea' (9).

3.3.2 *(the) Nature, (the) mountains*

Student translations:

the wonders of the Nature; those of you who enjoy walking in mountains

English dictionaries and grammars include sequences such as *the wonders of Nature* – without the definite article – and *those of you who enjoy walking in the mountains* – with the definite article – yet both of these represent usage which appears to fall outside the general rules of English grammar. In the sense of 'natural world', logic and consistency suggest that *the Nature* would be better, along the lines of other instances of huge natural phenomena such

as *the environment, the countryside, the sea, the world, the universe*, and even *the great outdoors*, where the definite article is used both when the reference is general (*please respect the environment*) and when it is more specific (*the environment of the eastern Baltic sea*). For this reason, it is perhaps not surprising that trainee translators recurrently adopt the article in sequences such as *the wonders of the Nature*, though corpus searches in the *British National Corpus and British Web 2007* establish that this type of sequence is barely attested.

The case of *those of you who enjoy walking in the mountains* also entails what is apparently anomalous usage relating to the definite article. The plural use of a countable noun with preceding *the* usually denotes a specific reference (*the kids in my class are boisterous, the computers we bought are defective*), but in fact *the mountains* is used both specifically (*we were hiking in the mountains to the north of Lake Maggiore*) and generally (*since she was a girl she's always loved walking in the mountains*), a state of affairs similar to *the hills* and *the woods*, which also tend to be accompanied by the definite article even when the sense is more general.

Indeed, in the *British National Corpus* there is only a handful of occurrences of *in mountains* (9), *into mountains* (4) and *to mountains* (11, though 3 are used in the sense of *a large pile of*) while there are 245 occurrences of *in the mountains*, 75 of *into the mountains* and 52 of *to the mountains*, a substantial percentage of which seem to be references of a general nature. In the *British Web 2007* the proportional frequencies of these sequences are in line with those of the *British National Corpus*, but trainees may be influenced by the raw numbers of examples without the article in the larger *British Web 2007*: 181 cases of *in mountains*, 36 of *into mountains* and 138 of *to mountains*.

It can certainly be hypothesised that the perennial difficulty Italian students experience with these phrases links with the fact that in Italian *natura* is preceded by the definite article (*la natura*) when the meaning corresponds to the natural world, and that *in/into/to the mountains* as a rule corresponds to *in montagna* in Italian, which has no article. At the same time, it seems just as legitimate to hypothesise that they are led astray by the anomalous local grammar of both *nature* and *mountains*.

3.3.3 Summing up

The situation regarding *view on the sea* etc. seems uncontroversial, but the presence / absence of the definite article in *the wonders of the Nature* and *those of you who enjoy walking in mountains* is less straightforward. Some evaluators will simply mark these as wrong, firstly because the sequences seem unidiomatic, but secondly because they appear to stem from ignorance, i.e., the student in question may have simply produced a clumsy and fairly literal rendering of *la natura* and *in montagna*. Yet this may not be the case at all: in the first instance the student may have made an intelligent analogy with *the countryside, the world*, etc., while in the second instance the student may have made the coherent assumption that *in mountains* should not include the definite article because the reference in question is general, i.e., there

is no reference to specific mountains. In a pedagogical situation, this might constitute grounds for showing a degree of leniency towards *the wonders of the Nature*, and perhaps even greater leniency – once the trainer ascertains that *in mountains* turns up 181 times in the *British Web 2007* – towards *those of you who enjoy walking in mountains*. It goes without saying that vastly different opinions will be held from one trainer to the next.

4. Other language features

Clearly the features put under the microscope above are merely a selection of those produced by Italian trainee translators in tourist texts, since many others could be discussed. On a lexical level, it is worth noting the persistent use of the verb *valorise* in the sense of *give or ascribe value or validity* to (this is the *Macmillan Dictionary* definition but interestingly this verb is not listed in the *Longman Dictionary of Contemporary English* or the *Oxford Advanced Learner's Dictionary*), and on a collocational level the overuse of combinations such as *valorise the area / territory / zone* – according to the *British National Corpus* and *British Web 2007* this verb normally combines with abstract nouns such as *notion, individuality* and *autonomy*.

Also recurrent in trainee translations is the use of Roman numerals with reference to centuries (*the XIV century, the XVIII century*), unusual in native-speaker English. Consider the following outcomes in the *British National Corpus* and *British Web 2007*:

British National Corpus	
fourteenth century	345
fourteenth-century	95
14th century	102
14th-century	28
XIV century	0
XIV-century	0

British Web 2007	
fourteenth century	1,819
fourteenth-century	334
14th century	4,472
14th-century	348
XIV century	18
XIV-century	0

Note that in the exclusively native-speaker corpus *fourteenth century* is much more common than *14th century*, whereas the reverse is true in the *British Web 2007*. As regards the formula *XIV century / XIV-century* with Roman numerals, there are a few attestations in the *British Web 2007* but none at all in the *British National Corpus*. This should probably be considered a mistake, especially since much of the intended international readership will struggle with Roman numerals. At the same time, regardless of possible incomprehension, Roman numerals tend to be adopted in English for popes (*Pope Pius XI, Pope John XXII*) and royalty (*Elizabeth II, Louis IV*), so it might seem churlish to judge *XIV century, XV century* etc. as being completely wrong.

5. Discussion

It emerges from the analysis so far that within the type of translation classroom scenario envisaged above, some solutions can be dealt with in fairly straightforward fashion while others are rather more contentious. The straightforward ones are those for which searches in dictionaries, grammars and corpora produce similar findings, for instance *only if, km far, at the feet of, view on the sea, nature*. It seems hardly worth stressing that the apparently more appropriate solutions are not necessarily the fruit of sounder logic. Nobody would argue, one imagines, that *the river is 50 metres wide* is somehow more logical than *the next town is 20 km far*, that *at the foot of the mountain* makes inherently more sense than *at the feet of the mountain*, that *the 20th century* is more logical than *the XX century*, or that the inversion of subject and verb after *only if* can be successfully rationalized. Frustrating, perhaps, but students can at least draw consolation from the fact that in these cases consistent answers are to be found if good searches are conducted in language resources.

The more contentious solutions are those for which language resources provide contrasting information. This may be from one dictionary to the next, for instance the plural *evidences* is included the *Oxford Advanced Learner's Dictionary*, whereas the *Macmillan Dictionary* and the *Longman Dictionary of Contemporary English* discountenance the plural use of *evidence*; similarly, the verb *valorise* (in reality *valorize*) is present in the *Macmillan* but not in the *Oxford* and the *Longman*. Information may also be in conflict (i) between dictionaries and corpora, for example the countable use of *scenery* is absent from dictionaries but moderately recurrent in the *British Web 2007*, and (ii) from corpus to corpus, as described above for the adjective *whole* + uncountable noun, *in the last years, the world war I, in mountains*, and for the countable usage of traditionally uncountable nouns.

It was also pointed out that, by an irony, some examples of questionable usage (for instance *only if* + SV in the main clause, *in mountains, the wonders of the nature*) are probably clearer for most of the intended readers than the 'correct' versions (*only if* + VS in the main clause, *in the mountains, the wonders of nature*). Where does this leave the translator trainer

within the framework outlined at the beginning of this paper (tourist texts from Italian to English where English is the students' L2; international readership mostly comprised of non-native speakers of English)?

As regards the cases I have described as 'straightforward' above, it seems to me that the trainer has little choice but to deem them inappropriate. I write this with some reluctance, as I am heartily sick of correcting *only if* + SV, *20 km far, view on the sea* etc., firstly because over the years I have done it so often, secondly because they are semantically transparent, thirdly because in some of these cases most of the target readers would understand the 'inappropriate' option better, and fourthly because such usage now appears to be part of global English – for instance in the gigantic *English Web 2020* corpus (*enTenTen20*), *km/s far* occurs 2,645 times, the sequence *view on the sea* occurs 333 times, and even *at the feet of the mountain/s* has 66 occurrences. Not to mention the results retrieved by Google – suffice it to say that the combinations *km / kms / kilometres / kilometers far* generate around two and a half million hits. The crucial point here is that if a trainer decides to accept sequences because they turn up very frequently in enormous web corpora or on Google, then s/he will be in the ludicrous situation of having to accept virtually everything (see Stewart 2013, 228-229). Consequently, this cannot be a viable methodology.

With regard to the cases I have described above as 'contentious' (adjective *whole* + uncountable noun, *in the last years, the world war I, in mountains, valorise*, the countable usage of traditionally uncountable nouns), in my view the situation is more complex. They lie for the most part outside resources based on native-speaker input (*Oxford Advanced Learner's Dictionary, British National Corpus* etc.), but once we begin to move beyond those exclusively native-speaker confines, for example into the uk-domain *British Web 2007*, the situation changes significantly. It is at this point, sadly, that evaluation risks turning into a lottery.

Some trainers will reject these contentious cases outright because they are not sufficiently backed up by native-speaker usage. Other trainers will half-accept them – treating them as minor errors – because though mostly absent from exclusively native sources they do turn up in uk-domain sources. And other trainers will accept them totally for an amalgamation of different reasons: they don't sound too bad, they are perfectly comprehensible (*the whole Europe*), they are very close to native usage (the structure of *the world war I* is analogous to the unobjectionable *the first world war*, and *in mountains* respects native-English norms), they fill a gap in native English (there appears to be no exact synonym of *valorise* in English), they can be more flexible than some 100% native counterparts (*in the last years* is temporally more flexible than the uncontroversially native *in the last few years*, which appears to restrict the temporal reference to the last 3-4 years), while the countable usage of traditionally uncountable nouns is part of a massive ongoing process in English and therefore a tolerant attitude is defensible. For this reason, it is advisable for trainers to clarify which parameters they prioritise right from the outset, though of course the choice of parameters will be highly subjective

and will thus differ markedly from one translation teacher to another.

6. Conclusions

This article is concerned with the enormous repercussions that the constantly evolving status of the English language can have upon the way trainers evaluate student translations. In the past I have advocated – in a European context – the use of British English for the translation scenario described in this paper, mostly because it offers a well-documented framework within which to operate, notwithstanding two main reservations: firstly, the patent incongruity of adopting a single variety of English for a global readership, and secondly, the perhaps not so patent incongruity of adopting a native variety of English for a predominantly non-native English readership.

The first of these reservations can be overcome – in Europe but also in many other parts of the world British English is a variety which non-native English speakers have studied from a young age. This, combined with the fact that it is flanked by a wealth of didactic materials, make it a good candidate for the translation classroom. In these respects it has a major advantage over ELF, which has not been taught in European schools and which at the present time still has very limited language resources to support it. The second of the reservations above is, in my view, a far greater obstacle. Why insist on native parameters when both translators and target are prevalently non-native? Why insist on native fluency when so many people around the world communicate successfully with non-native fluency? Why exclude the transparent and inoffensive *the whole Europe, the world war I, in the last years* in the sense of *in recent years* etc. simply because native speakers of English are not likely to produce them?

A cogent modus operandi could be that of allowing non-native sounding usage found with at least moderate frequency (though this too is subjective) in a uk-domain corpus such as the *British Web 2007*. This would at least represent a compromise between an exclusively native-speaker corpus such as the *British National Corpus* and a massive web-based corpus such as the *English Web 2020*. Of course, one could take the view that, within reason, it does not matter which variety of English or which language resources are prioritised, the crucial criterion is that trainer and trainee are clear about which have been chosen. I have subscribed to this position in the past, but the danger is that of requiring trainees to become mere information managers without any real connection to the current world of vocational translation. Truth be told, the situation is something of a muddle, with very different criteria being adopted from one translation classroom to the next. The muddle will presumably persist until such time as the transition towards a truly global language is more or less complete, with most of the world having expertise in a common language. If that common language is English, then this paper will – fortunately – be little more than a quaint historical relic.

Notes

[1] The *British Web 2007*, also known as *ukWac*, is a web-derived corpus assembled in 2007 containing over 1 billion 300 million words from websites in the uk Internet domain. It is a general-purpose corpus with a broad range of text types. The *British National Corpus* contains approximately 100 million words of British English from the late twentieth century. It too is a general-purpose corpus offering a broad range of text types. It contains 90% written texts and 10% spoken. The two corpora have been consulted via *The Sketch Engine*, a corpus manager and analysis software created by Lexical Computing Ltd in 2003, now with over 500 corpora in more than 90 languages. See https://www.sketchengine.eu (Last visited July 15 2021) for further details.

[2] One cannot exclude exceptional cases, for example *there is a wonderful view on the mountain* could in theory imply that the viewer is or was standing on the mountain (i.e., *the view is wonderful when you're actually on the mountain*), but I have not found any occurrences with this meaning in the two corpora, perhaps because in this type of sequence *from the (top of the) mountain* sounds more natural.

References

Anderman, Gunilla, and Margaret Rogers. 2005. "English in Europe: For Better, For Worse?." In *In and Out of English, For Better, For Worse?*, ed. by Gunilla Anderman, and Margaret Rogers, 1-26. Clevedon: Multilingual Matters.

Anderman, Gunilla, and Margaret Rogers. eds. 2005. *In and Out of English, For Better, For Worse?*. Clevedon: Multilingual Matters.

Buckledee, Steve. 2010. "Global English and ELT Coursebooks." In *EIL, ELF, Global English: Teaching and Learning Issues*, ed. by Cesare Gagliardi, and Alan Maley, 141-151. Bern: Peter Lang.

Bowles, Hugo, and Alessia Cogo. eds. 2015. *International Perspectives on English as a Lingua Franca: Pedagogical Insights*. Houndmills: Palgrave Macmillan.

Crystal, David. 2007. *English as a Global Language*. Cambridge: Cambridge University Press.

Gagliardi, Cesare, and Alan Maley. eds. 2010. *EIL, ELF, Global English: Teaching and Learning Issues*. Bern: Peter Lang.

Ife, Anne 2005. "Intercultural Dialogue: The Challenge of Communicating across Language Boundaries." In *In and Out of English, For Better, For Worse?*, ed. by Gunilla Anderman, and Margaret Rogers, 286-298. Clevedon: Multilingual Matters.

Jenkins, Jennifer. 2007. *English as a Lingua Franca: Attitude and Identity*. Oxford: Oxford University Press.

Jenkins, Jennifer, Will Baker, and Martin Dewey. eds. 2018. *The Routledge Handbook of English as a Lingua Franca*. London and New York: Routledge.

Maley, Alan. 2010. "The Reality of EIL and the Myth of ELF." In *EIL, ELF, Global English:*

Teaching and Learning Issues, ed. by Cesare Gagliardi, and Alan Maley, 25-44. Bern: Peter Lang.

Mossop, Brian. 2006. "Has Computerization Changed Translation?". *Meta: Journal des Traducteurs/Meta: Translators' Journal* 51 (4): 787-805.

Seidlhofer, Barbara. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Stewart, Dominic. 2008. "Vocational Translation Training into a Foreign Language." *inTRAlinea* 10. https://www.intralinea.org/index.php/specials/article/1646

Stewart, Dominic. 2013. "From Pro Loco to Pro Globo: Translating into English for an International Readership." *The Interpreter and Translator Trainer* 7 (2): 217-234.

Taviano, Stefania. 2010. *Translating English as a Lingua Franca*. Milano: Mondadori Education.

Taviano, Stefania. 2013. "English as a Lingua Franca and Translation." *The Interpreter and Translator Trainer* 7 (2): 155-167.

Dictionaries consulted

Longman Dictionary of Contemporary English Online.
https://www.ldoceonline.com. Accessed July 15 2021.

Macmillan Dictionary Online.
https://www.macmillandictionary.com. Accessed July 15 2021.

Oxford Advanced Learner's Dictionary Online.
https://www.oxfordlearnersdictionaries.com. Accessed July 15 2021.

Revisiting the (Overlooked) Landscape of CTIS – A Review of *CTS Spring-cleaning: A Critical Reflection*

Cui Xu

Address: Department of Foreign Languages, Beijing Institute of Technology, Beijing, China
E-mail: xcui0414@126.com
Correspondence: Cui Xu

María Calzada Pérez and Sara Laviosa (eds.). *CTS Spring-cleaning: A Critical Reflection / Reflexión Crítica En Los Estudios De Traducción.* Special Issue, *MonTI* 13(1). 2021. pp. 334. ISSN-e 1989-9335. ISSN 1889-4178. Spain: Publicacions de la Universitat d'Alacant.

1. Background

Since the introduction of corpus linguistics to translation and interpreting studies (TIS) in the 1990s (Baker 1993; Shlesinger 1998), great advances have been made in corpus-based translation and interpreting studies (CTIS), as evidenced by an overwhelming amount of journal articles and book publications in the field, notably De Sutter, Lefer and Delaere (2017), Fantinuoli and Zanettin (2015), Ji, Oakes, Li and Hareide (2017), Russo, Bendazzoli and Defrancq (2018), as well as Vandevoorde, Daems and Defrancq (2020), addressing various topics of corpus construction, recurrent features of translational (including interpreted) language, contrastive analysis, translator's style, and register variation. These vigorous efforts help accelerate the maturity of CTIS as a "theoretically robust and methodologically sound" (69) sub-discipline of TIS. However, at the turn of a new decade, during which time more sophisticated corpus tools are coming into existence and new perspectives are being adopted in CTIS, it seems to be the right time to pause a while and reflect upon the gains and losses over the past few decades in CTIS, which is exactly the purpose of the book under review.

CTS Spring-cleaning: A Critical Reflection, co-edited by María Calzada Pérez and Sara Laviosa and published by Publicacions de la Universitat d'Alacant in 2021, as a Special Issue of *MonTI*, is the first of its kind to practice self-reflection in the field of CTIS. As stated by the authors, the book aims "to encourage CTIS practitioners to pause and look around; to explore dusty corners and blind spots; to fight partiality, while injecting doses of innovation into our work". All in all, it aims "to boost critical thinking, (self-)awareness and (self)reflexivity,

without renouncing to our past" (23).

Overall, this edited volume is composed of 11 articles organized into four parts. The first two articles are the same introductory article written in English and Spanish, respectively. It first gives theoretical reflections on previous studies in CTIS, before moving on to review the empirical findings of relevant studies. Afterwards, the editors, i.e. María Calzada Pérez and Sara Laviosa, highlight the need to pause and look around the overlooked landscape of CTIS.

After the introduction, the topics in this volume focus on four key themes: "Translation Features as a Starting Point", "Neglected and Overlooked Areas of Study", "Researching Original and Translated Communication Under New Conditions", and "Self-reflexivity". Most of the contributions are dedicated to written and audiovisual modes of translation, except for the one by Marta Kajzer-Wietrzny and Łukasz Grabowski, which focuses on simultaneous interpreting based on an intermodal corpus, i.e., European Parliament Translation and Interpreting Corpus or EPTIC. Besides, the lion's share of this book is devoted to translations in the Spanish context.

2. The overlooked landscape of CTIS

Corpus-based translation and interpreting studies, as reviewed by María Calzada Pérez and Sara Laviosa in their introductory article, has paid excessive attention to the recurrent features of translation, or the quest for "translation universals" (Baker 1993). This is not surprising, as "the fundamental goal of any science" is to "look for regularities, generalities, [and] patterns" (Chesterman 2017, 139). To tighten its unbreakable bondage with this area of research, the current volume starts with the topic of translation features, before setting foot in the overlooked landscape of CTIS.

2.1 Translation features as a starting point

The selected paper in this part revisits one of the most widely discussed translation universals, namely, explicitation, and its under-researched opposite tendency, implicitation, in translated medical texts. Motivated by their previous comparable corpus study on explicitation (Jiménez-Crespo and Tercedor 2017), Jiménez-Crespo and Maribel Tercedor Sánchez carried out a follow-up study to test whether the higher explicitation ratios reported in their previous study are due to 1) cross-linguistic interference or 2) the translational tendency to explicitate. To that end, they utilized a parallel corpus of English to Spanish translations with a focus on Latin-Greek (LG) terms, which allowed them to compare directly source texts with target/translated texts. Overall, they found sufficient support for the cross-linguistic interference (CLI) hypothesis (Kruger 2019), meaning that explicitation of translated LG terms (compared to non-translated LG terms) are due to the replication of source language structures, regardless of linguistic differences between English and Spanish in LG term usage. Moreover, implici-

tation was found "even when it could be possible" (84), which may suggest a translational tendency to explicitate, lending "a more lukewarm support for the risk aversion hypothesis" (85). The authors concluded by emphasizing the suitability and necessity of a combination of comparable and parallel corpus methodologies in future studies.

2.2 Neglected and overlooked areas of study

This section of the volume aims "to explore dusty corners and blind spots" (12) in CTIS. It includes four articles, dealing with such topics as subtitling (Blanca Arias-Badia), translation of travel journalism (David Finbar Brett, Barbara Loranc-Paszylk and Antonio Pinna), constrained communication (Kajzer-Wietrzny and Łukasz Grabowski), and operatic audio description (Irene Hermosa-Ramírez), topics that have received far less attention from CTIS scholars.

The article by Blanca Arias-Badia provides a methodological reflection upon Corpus Pattern Analysis (CPA), a methodological framework adapted from the field of lexical analysis to determine patterns (normal or creative) of word usage in audiovisual translations. Based on a parallel (English-Spanish) corpus composed of three TV series, i.e. *Castle (2009), Dexter (2006)*, and *The Mentalist (2008)*, with a focus on anomalous collocates and lexical creativity (or lexical exploitation) in their subtitles, the author summarized both merits and demerits of the adapted CPA framework, offering its readers much food for thought in future studies.

The next article, co-authored by David Finbar Brett, Barbara Loranc-Paszylk and Antonio Pinna, looks into adjective/noun collocations in travel journalism (a much-overlooked area of research) through the lens of corpus linguistics. Taking advantage of a multilingual corpus comprising three comparable sub-corpora of travel reportage in English (from *The Guardian*), Italian (from *La Repubblica*) and Polish (from *Gazeta*), they attempted to address several issues concerning adjective/noun collocations in the three languages, such as the differences and similarities in the frequencies of adjective/noun collocations, connectivity, syntactic variability and collocations in selected themes (116). Their ultimate aim, as stated by the authors, was to "[garner] information useful to translators [and practitioners]" (140). This research gives full play to the role of corpus linguistics techniques. It demonstrates the power of statistical sophistication (such as programming in Perl) as well as other tools outside corpus linguistics (such as Gephi from Social Networks Analysis) in facilitating corpus analysis. Besides, this study also shows how research findings can be fed into translator training and teaching in the future.

Exploring another form of collocations (i.e. bigrams) from the perspective of constrained communication (focusing on translation, interpreting and L2), Kajzer-Wietrzny and Łukasz Grabowski set out to explore the factors/predictors (i.e. language variety, mode of delivery, delivery rate and text ID) for the degree of formulaicity in constrained communication based on the Polish-English subcorpora of the EPTIC intermodal corpus. Backed by robust statisti-

cal modelling, namely, a mixed-effects model using Poisson regression, their study revealed several enlightening patterns. To begin with, be they written or spoken, constrained varieties use more frequent bigram types than non-constrained (i.e. native) varieties, demonstrating similarities shared among constrained languages. Nonetheless, translated variety (including translation and interpreting) exerts a greater effect than non-translated constrained communication (i.e. L2 spoken and written language). Secondly, be it written or spoken, mode of delivery (i.e. whether the source text is read-out or delivered impromptu) is closely associated with the frequency of bigrams. Specifically, more frequent bigram types are observed in translated variety when the source text is delivered impromptu than read-out, and the trend is more significant in the written register, which points to an equalizing effect in interpreting (see also Shlesinger and Ordan 2012). Moreover, delivery rate also has an influence over the use of bigram types, but the effect is not statistically significant. Overall, these variations can be accounted for by random variables rather than fixed variables.

The last article by Irene Hermosa-Ramírez in this part addresses a niche topic that has rarely been investigated in CTS, namely, operatic audio description (AD), which falls into the category of intersemiotic translation according to Jakobson (1959). Based on a multimodal corpus consisting of opera AD scripts from two famous opera houses, i.e. the Teatro Real in Madrid (the Spanish corpus) and the Liceu in Barcelona (the Catalan corpus), and with the aid of Sketch Engine, the author investigated the lexico-grammatical features of the two scripts, and analyzed their operatic semantic meanings drawing from Rędzioch-Korkuz's proposed framework (2016) and the TRACCE narratology tagset (187, 205). Her study revealed that the two AD scripts resemble each other in a number of lexico-grammatical features, such as nouns, verbs, PoS distribution, type-token ratio (TTR) and standardized type-token ratio (STTR). However, discrepancies also exist in terms of mean sentence length, which can be attributed to the different strategies adopted in the two opera houses. With respect to the semiotic dimension, similarities have been identified in sign saliency in the two AD scripts, in which character identification with proper and common nouns are particularly salient. However, the two diverge in scenography and surtitles. The author concluded by highlighting the challenges of using annotated multimodal corpora in this study.

2.3 Researching original and translated communication produced under new conditions

This part includes two articles reporting corpus findings against the background of localization (Laura Mejías-Climent) and an increasingly demanding whilst decreasingly cost-effective translation market (Leticia Moreno-Pérez and Belén López-Arroyo).

The first article by Laura Mejías-Climent, written in Spanish, ventures into the game field with a focus on localization (specifically, dubbing synchronies) of three action-adventure video games of an interactive genre from English to Castilian Spanish based on a multimodal

corpus (MMC). Resorting to data triangulation, which combines quantitative (types of synchronies/adjustment) and qualitative data (semi-structured interviews), and putting the analysis in the AVT paradigm¸ the authors identified a relationship between certain game situations (i.e. tasks, cinematics, game action, and dialogues) and types of synchronies/adjustment (e.g. free, lip-sync), which was further corroborated by semi-structured interviews. Specifically, they found that tasks always opt for free-sync (unrestricted translation such as voice-over), while in cinematics there seems to be a clear preference for lip-sync (238-239). Curiously though, they also found that sound technicians instead of translators shouldered more responsibilities of dubbing audio waves to the original, which is very different from film dubbing (241-242). They concluded their article by stating some of the limitations of their study, including a potential mismatch between the original version and the translated version, and the exclusion of other game genres. For addressing these issues, they also called for more cooperation with the translation industry.

The second article by Leticia Moreno-Pérez and Belén López-Arroyo illustrates how (atypical) corpus tools can help translators adapt to the demands of the market. To make their points more explicit, they first studied the reality of the translation market – increasingly demanding whilst decreasingly cost-effective – to understand the pressing needs of translators. They then moved on to introduce some of the available corpus resources that can come to the translators' rescue, such as writing assistants, templates, and writing generators (259), as well as some typical corpus resources such as lexicographical resources, translation memories, machine translation, online corpora, and web crawlers (257-258). Afterwards, they introduced the ACTRES oenology generator, which is based on comparable corpora of wine tasting notes in Spanish and English, respectively, to illustrate step by step how the writing generator can help improve translators' efficiency in terms of costs, time, and quality. Ending on a positive note, they pointed out that given the reality of the current translation market, the use of writing generators may actually make a difference.

2.4 Self-reflexivity

Corpus-based translation and interpreting studies is never short of empirical studies. However, this is not the case for reflexive works, which has made the last two articles in this volume particularly thought-provoking.

The article "Autocrítica de publicaciones previas basadas en corpus: Análisis DAFO", co-authored by Alexandra Santamaría Urbieta and Elena Alcalde Peñalver, serves as the very example of self-reflexivity by reflecting upon four previous publications written by the authors using SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis. Simply put, they analyzed the strengths, weaknesses, opportunities, and threats (obstacles to publication) of each of the four publications so as to reflect upon the way forward in corpus-based translation studies. However, the readers would get more inspirations if the authors could provide a

more detailed SWOT analysis specifying the corpus tools used, the quantitative or qualitative analysis carried out, as well as the main findings.

The last article co-authored by Jan Buts and Henry Jones promotes deeper reflection on the role of mediality, a dusty corner in CTS research, with reference to the KWIC display. In their view, a medium, which is "a channel of communication", can shape a discipline, as is argued to be the case with corpus-based translation studies (305). CTS is essentially a field of study where the technological and theoretical boundaries converge, and because of that, "what suggests itself as a theoretical discrepancy is in fact a technological one" (309). They illustrated their points with reference to the role of the KWIC concordancer alongside two visualization tools (i.e. the Metafacet tool and the Mosaic tool) based on the two English versions (i.e. MacFarlane's and Samuel Moore's 1888 version) of the *Communist Manifesto* in the medial environment of the Genealogies of Knowledge (GoK) software. Their findings indicated that the tools they use (i.e. the medium) can heavily influence their interpretation, since the medium is "not a neutral tool of representation" (306). Given the fact that different tools may have prioritized different theories and principles, they called for the triangulation of multiple software tools during data interpretation.

This article is particularly thought-provoking in the current context, where various corpus tools are easily available and CTIS studies are often carried out utilizing these tools which in essence prioritize different principles. The major takeaway from this article is that, when we make an interpretation in a customized medial environment (such as an *ad hoc* corpus), we should also evaluate critically what this environment offers to us.

3. The way forward

Overall, this edited volume offers refreshing perspectives in the field of corpus-based translation and interpreting studies. It examines closely the "dusty corners" and overlooked areas from theoretical, empirical, and methodological perspectives, covering such topics as subtitling, travel journalism, localization, interpreting, and operatic audio description. It also includes detailed references that are often neglected in CTIS, guiding readers to further explore the under-explored fields. Particularly, its inclusion of the study on operatic audio description (by Irene Hermosa-Ramírez), a form of intersemiotic translation, as compared to the 'taken-for-granted' intralingual translation or translation proper (Jakobson 1959), is extremely refreshing and fascinating.

Based on this enlightening volume, I see several ways forward in corpus-based translation and interpreting studies. First of all, future CTIS research needs to "walk out of their comfort zone" (e.g. literary translation) by venturing into some less-trodden fields, such as localization of different genres of game videos, theatre translation, and voice-overs in TVs and movies. This is actually becoming the trend, as an increasing number of studies are shifting

their focus from the canonical literary translation to other translation modalities, such as audiovisual translation (e.g. dubbing, subtitling). Closely related to the first point, future CTIS studies need to continue to "look over the disciplinary fence" (De Sutter and Lefer 2020, 2) by expanding their research topics. For example, studies on audiovisual translations will inevitably involve other disciplinary studies such as Television Studies, Film Studies, Opera Studies, etc., as has been the case in the current volume. Thirdly, answering the call by De Sutter and Lefer (2020), more sophisticated and robust statistical analyses will be a prominent trend in future studies. This has already been the case, exemplified not only by studies included in this volume, but also by an increasing number of recent publications (e.g. De Sutter and Vermeire 2020; Kajzer-Wietrzny and Ivaska 2020; Kruger 2019; Kruger and De Sutter 2018), featuring R or Python programming. More robust and detailed statistical testing can allow and encourage more fine-grained replication studies to be conducted. In addition, data triangulation seems to be a proper choice in future studies, as "they allow not only the researcher but also the professional translator to look into the data from many different windows" (257). Last but not least, future corpus-based studies should continue to narrow the gap between research/academia and practice by feeding research findings into translational practice, and vice versa, to address the long-held criticism held by practitioners over the usefulness of research.

Nonetheless, several limitations should be pointed out. To begin with, while this volume aims to pause and reflect upon both translation and interpreting studies utilizing corpus tools, it is skewed towards the corpus-based translation studies (CTS) relatively at the expense of corpus-based interpreting studies (CIS). Although it has been a well-known fact that interpreting studies (IS) has been, in many ways, lagging behind compared to translation studies (TS), researchers have never ceased their efforts to reunite the two "sister disciplines" (Defrancq, Daems and Vandevoorde 2020). A coordinated progress in both lines of inquiry is, therefore, expected. Besides, the majority of this volume is dedicated to translations carried out in the Spanish context involving the Spanish/English language pair, albeit some focus on other European languages such as Italian and Polish. An inclusion of non-European languages such as Chinese, Japanese, or Afrikaans may represent better the overall landscape of CTIS.

Acknowledgment:
This work was supported by a grant from Beijing Social Science (19YYB011)

References

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies: Implications and Applications." In *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233-250. Amsterdam and Philadelphia: John Benjamins.

Chesterman, Andrew. 2017. *Reflections on Translation Theory*. Amsterdam and Philadelphia:

John Benjamins.

Defrancq, Bart, Joke Daems, and Lore Vandevoorde. 2020. “Reuniting the Sister Disciplines of Translation and Interpreting Studies.” In *New Empirical Perspectives on Translation and Interpreting*, ed. by Lore Vandevoorde, Joke Daems, and Bart Defrancq, 11-21. London and New York: Routledge.

De Sutter, Gert, Marie-Aude Lefer, and Isabelle Delaere. eds. 2017. *New Methodological and Theoretical Traditions*. Berlin and Boston: Walter de Gruyter GmbH.

De Sutter, Gert, and Marie-Aude Lefer. 2020. “On the Need for a New Research Agenda for Corpus-based Translation Studies: A Multi-methodological, Multifactorial and Interdisciplinary approach.” *Perspectives* 28 (1): 1-23.

De Sutter, Gert, and Eline Vermeire. 2020. “Grammatical Optionality in Translations: A Multifactorial Corpus Analysis of that/zero Alternation in English Using the MuPDAR Approach. In *New Empirical Perspectives on Translation and Interpreting*, ed. by Lore Vandevoorde, Joke Daems, and Bart Defrancq, 24-51. London and New York: Routledge.

Fantinuoli, Claudio, and Federico Zanettin. eds. 2015. *New Directions in Corpus-based Translation Studies*. Berlin: Language Science Press.

Jakobson, Roman. 1959. “On Linguistic Aspects of Translation”. In *On Translation*, ed. by Reuben Arthur Brower, 232–239. Cambridge: Harvard University Press.

Ji, Meng, Michael Oakes, Defeng Li, and Lidun Hareide. eds. 2017. *Corpus Methodologies Explained. An Empirical Approach to Translation Studies*. London and New York: Routledge.

Jiménez-Crespo, Miguel A., and Maribel Tercedor. 2017. “Lexical Variation, Register and Explicitation in Medical Translation: A Comparable Corpus Study of Medical Terminology in US Websites Translated into Spanish.” *TIS: Translation and Interpreting Studies* 12 (3): 405-426.

Kajzer-Wietrzny, Marta, and Ilmari Ivaska. 2020. “A Multivariate Approach to Lexical Diversity in Constrained Language.” *Across Languages and Cultures* 21 (2): 169-194.

Kruger, Haidee. 2019. “That Again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English.” *Across Languages and Cultures* 20 (1): 1-33.

Kruger, Haidee, and Gert De Sutter. 2018. “Alternations in Contact and Non-contact Varieties: Reconceptualising That-omission in Translated and Non-translated English Using the MuPDAR Approach.” *Translation, Cognition & Behavior* 1 (2): 251-290.

Russo, Mariachiara, Claudio Bendazzoli, and Bartt Defrancq. eds. 2018. *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer.

Shlesinger, Miriam. 1998. “Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies.” Meta: Journal des Traducteurs/Meta: Translators’ Journal 43(4): 486-493.

Shlesinger, Miriam, and Noam Ordan. 2012. “More *Spoken* or More *Translated*?: Exploring

a Known Unknown of Simultaneous Interpreting." *Target* 24(1): 43-60.
Vandevoorde, Lore, Joke Daems, and Bart Defrancq. eds. 2020. *New Empirical Perspectives on Translation and Interpreting*. London and New York: Routledge.

稿約凡例 Guidelines for Contributors

《翻譯季刊》為香港翻譯學會之學報，歡迎中、英文來稿及翻譯作品 (請附原文及作者簡介)。有關翻譯作品及版權問題，請譯者自行處理。

一、稿件格式

1. 請以電郵傳送來稿之電腦檔案。

2. 來稿請附 200-300 字英文論文摘要一則，並請注明：(1) 作者姓名；(2) 任職機構；(3) 通訊地址／電話／傳真／電子郵件地址。

3. 來稿均交學者審評，作者應盡量避免在正文、注釋、頁眉等處提及個人身份，鳴謝等資料亦宜於刊登時方附上。

4. 來稿每篇以不少於八千字（約 16 頁）為宜。

二、標點符號

1. 書名及篇名分別用雙尖號（《》）和單尖號（〈〉），雙尖號或單尖號內之書名或篇名同。

2. “ ” 號用作一般引號；‘ ’號用作引號內之引號。

三、子 目各段落之大小標題，請依各級子目標明，次序如下：
一、／ A. ／ 1. ／ a. ／ (1) ／ (a)

四、專有名詞及引文

1. 正文中第一次出現之外文姓名或專有名詞譯名，請附原文全名。

2. 引用原文，連標點計，超出兩行者，請另行抄錄，每行入兩格；凡引原文一段以上者，除每行入兩格外，如第一段原引文為整段引錄，首行需入四格。

五、注 釋

1. 請用尾注: 凡屬出版資料者，請移放文末參考資料部份。號碼一律用阿拉伯數目字，並用 () 號括上；正文中之注釋號置於標點符號之後。

2. 參考資料: 文末所附之參考資料應包括：(1) 作者／編者／譯者；(2) 書名、文章題目；(3) 出版地；(4) 出版社；(5) 卷期／出版年月；(6) 頁碼等資料，務求詳盡。正文中用括號直接列出作者、年份及頁碼，不另作注。

六、版　權

來稿刊登後，版權歸出版者所有，任何轉載，均須出版者同意。

七、贈閱本

從 2009 年夏天開始，作者可於 EBSCO 資料庫下載已發表的論文。如有需要，亦可向編輯部申領贈閱本。

八、評　審

來稿經本學報編輯委員會審閱後，再以匿名方式送交專家評審，方決定是否採用。

九、來稿請寄： translationquarterly@gmail.com

Guidelines for Contributors

1. Translation Quarterly is a journal published by the Hong Kong Translation Society. Contributions, in either Chinese or English, should be original, hitherto unpublished, and not being considered for publication elsewhere. Once a submission is accepted, its copyright is transferred to the publisher. Translated articles should be submitted with a copy of the source text and a brief introduction to the source-text author. It is the translator's responsibility to obtain written permission to translate.

2. Abstracts in English of 200–300 words are required. Please attach one to the manuscript, together with your name, address, telephone and fax numbers and email address where applicable.

3. In addition to original articles and book reviews, review articles related to the evaluation or interpretation of a major substantive or methodological issue may also be submitted.

4. Endnotes should be kept to a minimum and typed single-spaced. Page references should be given in parentheses, with the page number(s) following the author's name and the year of publication. Manuscript styles should be consistent; authors are advised to consult earlier issues for proper formats.

5. Chinese names and book titles in the text should be romanised according to the "modified" Wade-Giles or the pinyin system, and then, where they first appear, followed immediately by the Chinese characters and translations. Translations of Chinese terms obvious to the readers (like wenxue), however, are not necessary.

6. There should be a separate reference section containing all the works referred to in the body of the article. Pertinent information should be given on the variety of editors available, as well as the date and place of publication, to facilitate use by the readers.

7. All contributions will be first reviewed by the Editorial Board members and then anonymously by referees for its suitability for publication in Translation Quarterly. Care

should be taken by authors to avoid identifying themselves. Submissions written in a language which is not the author's mother-tongue should perferably be checked by a native speaker before submission.

8. Electronic files of contributions should be submitted to translationquarterly@gmail.com.
9. Given the accessibility, from summer 2009, of the journal via the EBSCO database, authors will no longer receive complimentary copies unless special requests are made to the Chief Editors.

《翻譯季刊》徵求訂戶啟事

香港翻譯學會出版的《翻譯季刊》是探討翻譯理論與實踐的大型國際性學術刊物，創刊主編為劉靖之教授，榮譽主編為陳德鴻教授，現任主編為李德超博士。學術顧問委員會由多名國際著名翻譯理論家組成。資深學者，如瑞典諾貝爾獎評委馬悅然教授、美國學者奈達博士及英國翻譯家霍克思教授都曾為本刊撰稿。《翻譯季刊》發表中、英文稿件，論文摘要（英文）收入由英國曼徹斯特大學編輯的半年刊《翻譯學摘要》。欲訂購的單位或個人，請聯絡：

中文大學出版社

地 址：香港 新界 沙田, 香港中文大學, 中文大學出版社

電 話：+ 852 3943 9800

傳 真：+852 2603 7355

電 郵：cup-bus@cuhk.edu.hk

網 址：www.chineseupress.com

Subscribing to Translation Quarterly

Translation Quarterly is published by the Hong Kong Translation Society, and is a major international scholarly publication. Its founding Chief Editor is Professor Liu Ching-chih and its Honorary Chief Editor is Professor Leo Tak-hung Chan. The current Chief Editor is Dr. Dechao Li. The Academic Advisory Board of the journal is composed of numerous internationally renowned specialists in the translation studies field. The journal has previously included contributions from such distinguished scholars as the Swedish Nobel Prize committee judge Professor Göran Malmqvist, the American translation theorist Dr. Engene A. Nida, and the English translator Professor David Hawkes. *Translation Quarterly* publishes contributions in both Chinese and English, and English abstracts of its articles are included in *Translation Studies Abstracts*, edited by UMIST, UK. Institutions or individuals who wish to subscribe to the journal should contact:

The Chinese University Press

Address: The Chinese University Press, The Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong

Tel: +852 3943 9800

Fax: +852 2603 7355

Email: cup-bus@cuhk.edu.hk

Website: www.chineseupress.com

Translation Quarterly 翻譯季刊

Subscription Information

- Subscriptions are accepted for complete volumes only
- Rates are quoted for one complete volume, four issues per year
- Prepayment is required for all orders
- Orders may be made by check (Payable to **The Chinese University of Hong Kong**) in Hong Kong or US dollars, or by Visa, MasterCard or American Express in Hong Kong dollars
- Orders are regarded as firm and payments are not refundable
- Rates are subject to alteration without notice

- **Orders and requests for information should be directed to:**
  The Chinese University Press
  The Chinese University of Hong Kong
  Sha Tin, New Territories, Hong Kong
  Tel: +852 3943 9800
  Fax: +852 2603 7355
  E-mail: cup-bus@cuhk.edu.hk
  Web-site: www.chineseupress.com

TO: The Chinese University Press Fax: +852 2603 7355

Order Form

Please enter my subscription to
Translation Quarterly, beginning with *No.95 to No.101 (2021)*

Subscription and order	Rates
1 year	☐ HK$624 / US$80
2 years*	☐ HK$1,123 / US$144
3 years**	☐ HK$1,498 / US$192
Back issues (No.1 - No.98)	☐ HK$180 / US$23 each (Please list issue no._____, total_____issues.)

Please circle your choice.
Prices are at discount rate, delivery charge by surface post included.
* 10% discount.
** 20% discount.

☐ Attached is a check in HK$ / US$* _____ made payable to "**The Chinese University of Hong Kong**". (*circle where appropriate)

☐ Please debit my credit card account HK$_____. (Please convert at US$1 = HK$7.8)

I would like to pay my order(s) by: ☐ AMEX ☐ VISA ☐ MASTER CARD

Card No. (including the 3-digit security code): _____

Expiry Date: _____

Cardholder's Name: _____

Cardholder's Signature: _____

Please send my journal to:

Name:_____

Address:_____

Telephone:_____Fax:_____E-mail: _____

Ref: 20210608

中文大學出版社
THE CHINESE UNIVERSITY PRESS
www.chineseupress.com
HONG KONG, CHINA

The Chinese University Press
The Chinese University of Hong Kong, Sha Tin, Hong Kong
Tel.: +852 3943 9800 Fax: +852 2603 7355 E-mail: cup-bus@cuhk.edu.hk
Web-site: www.chineseupress.com