Meta
Journal des traducteurs
Translators' Journal

META

Introduction: Quantitative methods in health, environmental, and theoretical translation research

Michael Oakes and Meng Ji

Volume 67, Number 1, April–May 2022

Pour de nouvelles méthodes en traductologie quantitative
Exploring New Methods in Quantitative Translation Studies

URI: https://id.erudit.org/iderudit/1092188ar
DOI: https://doi.org/10.7202/1092188ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this document

Oakes, M. & Ji, M. (2022). Introduction: Quantitative methods in health, environmental, and theoretical translation research. *Meta*, *67*(1), 5–17. https://doi.org/10.7202/1092188ar

Introduction: Quantitative methods in health, environmental, and theoretical translation research

MICHAEL OAKES
University of Wolverhampton, Wolverhampton, United Kingdom
Michael.oakes@wlv.ac.uk

MENG JI
University of Sidney, Sidney, Australia
christine.ji@sydney.edu.au

1. Context

Translation studies is a field of interdisciplinary research, and one of the factors that has driven its rapid development in recent times is methodological innovation. Translation scholars are one of the most dynamic research communities working across the boundaries of the arts and humanities, social and natural sciences. From comparative literature, bilingual and multilingual education to textual statistics, technology localisation, and machine learning modelling of large multilingual translation databases, for decades, we have been working passionately, tirelessly to advance the understanding of cross-cultural, cross-lingual translation. Despite the availability of constantly improving automatic machine translation technologies, translation studies, which explore the underlying principles, methods, and mechanisms of human translation activities, have strived around the world. This reflects increasing demands from different cultures, societies, and communities for high quality human translations and human-centred translations, which cannot be replaced by machine translation algorithms. Translation is never a straightforward activity. It centrally reflects the complex, subtle, and context-dependent nature of human communication. Existing translation theories tend to be operational at the macro level, exploring the social and cultural impact on the selection and use of certain translation strategies conceptualised as language-independent norms. This theoretical approach has proven effective when the reception of translation is a collective social behaviour. In recent studies, more importance is given to the impact of individual differences among readers on the design and development of translation resources, for example, for the purpose of public health education and communication. Translations which are adaptive to the varying reading habits and abilities of individuals tend to be successful with regards to their communicative effectiveness. From the communication of health risks to climate change, translation is playing an important role in reducing health and environmental inequalities, misunderstandings, and confusion in a world of uncertainties. New research questions that have emerged in our time include how to translate credible, critical information, for example, on health and environmental

issues more effectively to global readers. Research papers in this special issue illustrate that the development of multilingual resources for environmental communication represents another contribution that the translation community is making to the broader academy and to societies. For the purposes of health education and promotion, health translation requires higher understandability, readability, and accessibility. To translate effectively requires an in-depth understanding of the practical needs, reading habits, and reading abilities of the readers as individuals. This represents a missing link in many current translation practices. Even with machine translation, which exhibits increasing accuracy and fluency, few studies have addressed the pressing needs for human-centred translations. This special issue aims to foster scholarly debate around the value of new research evidence and the development of research methods to effectively process, analyse, and interpret translations. This is a continuation of the empirical translation studies envisaged by earlier scholars. The ongoing pandemic provides the social background for this issue on the data turn in translation studies, around which we discuss the development of integrated quantitative and qualitative approaches to address socially oriented research questions.

2. Understandability of translation

Three of the articles in this special issue investigate the question of translation understandability. The first is Silvia Rodriguez-Vasquez, Abigail Kaplan, Pierrette Bouillon, Cornelia Griebel, and Razieh Azari's contribution, entitled "La traduction automatique des textes faciles à lire et à comprendre: une étape comparative" [Machine translation of texts that are easy to read and understand: a comparative study]. The use of controlled languages (CL) has long been associated with machine translation (MT). Early MT systems such as TAUM-METEO, which translated weather reports between French and English, owed their success to the absence of ambiguity, as well as the restricted range of vocabulary and syntactic structures, in the specific sublanguage that they were translating. Over the last decade, many studies have looked at the influence of CL inputs on the quality of MT output, but this paper looks specifically at the influence of MT on the translation of texts specially written to be easy to read and understand by people with special communication needs. Other people who can benefit from Easy-to-Read (EtR) language are immigrants, asylum seekers, dyslexic people, people on the autistic disorder spectrum, people with mild intellectual disorders (the main group of beneficiaries), readers of public health information, learners of foreign languages, people who are deaf or hard of hearing, and the elderly. Even though EtR texts are not generally covered in the literature regarding CL, they can be considered a form of CL (Kuhn 2014). Techniques for producing EtR scripts in French have been developed by the Swiss Bureau fédéral de l'égalité pour les personnes handicapées [Federal bureau of equality for handicapped people] (BFEH).[1] Recommended techniques include the use of frequent words and short phrases, as well as language-independent norms for ideal page layout, such as the use of images and colour contrasts. Moreover, foreign words, contractions, abbreviations, and complex syntax should be avoided.

The work described in the article follows from Felici and Griebel (2019), who measured the quality of MT and the conformity of the output with the rules of Inclusion Europe[2] for the French-English language pair, but it goes further than this

exploratory study by looking at a broader range of languages and domains. The overall aim of Rodriguez-Vasquez, Kaplan, *et al*. is to determine if MT, when fed with EtR texts in the source language, is able to produce accessible texts in the target language. The research questions are a) what is the final quality of machine-translated EtR text, and b) does it preserve the rules of EtR? To answer these, the authors first count the number of errors then the number of times EtR rules are broken in the target text, according to the MT system used, the language pair, and the domain. The rules of EtR selected for the study are a subset of the those laid down by Inclusion Europe and Unapei,[3] such as: "Don't use difficult words. If you must use difficult words, you must explain them clearly"; "Use the same word to speak about the same thing throughout the document"; "Use simple punctuation"; and "Don't use long words. If you must use long words, separate them with a hyphen." To measure the quality of MT output, they use a typology of errors called the *DQF-MQM*.[4] Each annotation was done by two different annotators—all translation segments deemed incorrect contained at least one error marked by one of the annotators. Types of errors considered were mistranslations, lack of idiomaticity, and violations of free flow.

Three free, online MT systems are compared: the neural *DeepL* and *Google Translate*, as well as the hybrid neural and statistical *Yandex*. Their performance is evaluated in three different domains: administration, medicine, and politics. The evaluations are carried out for four language pairs: French to English, Persian, German, and Spanish. They found that *Yandex* produced the most errors in translation, while *Google* and *DeepL* were about equal. Administration proved to be the most difficult domain. The experiment worked less well with Spanish compared to the other languages, the best results being obtained for English. Similar results were obtained for all three MT systems, domains, and languages with regards to violations of the rules of EtR texts.

It is still not possible to fully translate EtR texts automatically, independently of MT system, domain, and language pair. Yet, in terms of language quality, *DeepL* performs best by a small margin. This confirms the results of numerous studies, which have shown that *DeepL* is in various ways the best MT system when using standard language. Spanish was the most problematic language, even though Spanish comes from the same language family as French. Overall, it was found that, although MT systems are not yet able to translate EfT well enough for practical use, the idea of using MT to translate EtR texts shows promise. In this article, the authors present (to our knowledge) the first study on the automatic translation of accessible texts, with different MT systems, domains, and language pairs. Their second contribution is the creation of a multilingual corpus, annotated according to the quality and accessibility of the translated segments, which will be of value for future research.

The second article that deals with translation understandability is Thomas François and Marie-Aude Lefer's "Revisiting simplification in corpus-based translation studies: insights from readability research," which focuses on the simplification and convergence hypotheses in translationese. The simplification hypothesis, first put forward by Laviosa (1998), states that translated texts are simpler than non-translated texts. She also proposed the convergence hypothesis: translated texts are more homogeneous than original texts, that is they show less variance, such as inter-speaker variation. This paper examines both hypotheses using measures derived from NLP-informed readability research, which allows the authors to compare original

French and translated French (from English) in the *Europarl*[5] corpus. They show that translated French is both lexically and syntactically simpler than original French, and that convergence occurs when less variation between different speakers is found in translated French. Previous empirical studies have produced mixed results; some have even found complexification in translated texts. François and Lefer aim to clarify the situation by using measures from readability research, such as readability formulas, which are designed to determine reading difficulty, often in terms of the average age of people first able to read a given text. Such measures might better capture simplification patterns in translation. Early measures included the number of syllables per 100 words and average sentence length. Most studies to date have looked at machine translations rather than texts translated by humans. Here the *Europarl* corpus is used, which contains human translations. The hypothesis is that speeches originally made in English but translated into French are simpler and display less variance than comparable speeches in their original French.

To test this, 19 measures were used, 15 of which were at the lexical level, such as the type-token ratio for lemmas and the type-token ratio normalized per hundred words. Two measures looked at syntactic complexity: average number of words per sentence and percentage of sentences longer than 30 words. The two measures that assessed discourse complexity were ratio of pronouns to proper names and ratio of pronouns to words in general. The chosen measures have been found in the past to be useful in readability research, and to be transparent, so that the findings would be meaningful in translation research. To measure these, the two subcorpora were part-of-speech tagged.

In the simplification analysis, the Wilcoxon statistical test was used to compare the values of each readability measure for translated and original texts, firstly by speaker and secondly by speech. In each case, the two discursive features showed an increase in complexity, but the vast majority of the other features showed that simplification had taken place during the translation process. For the experiments looking at convergence in translation between speakers, convergence was found for 11 features, and divergence only for the discursive feature of the pronoun to all nouns ratio. Thus, translation seems to smooth out inter-speaker differences at the lexical and semantic levels.

Another study which explores the impact on the difficulty of English source texts on the understandability of translations is the one by Diana Zaval-Rojas, Danielly Sorato, Lidun Hareide, and Knut Hofland on the multilingual translations of social surveys, "The multilingual corpus of survey questionnaires: a tool for refining survey translation." The authors describe the design and compilation of the first publicly available corpus of multi-lingual survey questionnaires, the *Multilingual Corpus of Survey Questionnaires* (MCSQ).[6] This contains both English source language questionnaires and the corresponding target translations in eight other languages and 29 of their varieties. The questionnaires come from the following surveys: *The European Social Survey*,[7] the *European Values Study*,[8] the *Survey of Health, Ageing and Retirement in Europe*,[9] and the *WageIndicator Survey*.[10] One advantage of the corpus is to enable analysis of sentences that are difficult to translate. The MCSQ is open access and open source, and consists of over 4 million words. The authors show that the MCSQ is a valuable resource, which has the potential to improve the translation of questionnaires.

A rigorous protocol for the translation of survey questionnaires is *TRAPD*, designed by Harkness (2003), where at least two people independently translate the original, these translations are reconciled, and an adjudicator makes the final decisions on different translation options. The questionnaire is pilot studied before being used in the final survey, and the whole process is documented. The *TRAPD* method roughly coincides with the "empirical turn" in translation studies. Hareide (2019) states that "this shift in translation studies was inspired by the paradigm change in linguistics from prescriptive to descriptive grammar, due to the corpus linguistic method." In the MCSQ corpus, the questionnaires are divided into various segments: introduction, instruction, request, and response. The corpus is part-of-speech tagged using language models learned by neural networks, and also tagged for named entities (real-world objects) such as locations and organizations, also using pre-trained neural models. The corpora are stored in electronic form by means of an entity-relation model. Details of how to obtain the corpus are given. Before entering the *TRAPD* process, the corpus allows examination of previous translations to review which sentences have been problematic to translate, and also to highlight examples of past success. In the translation step itself, the corpus allows examination of past translation variants. The adjudication team can benefit from a statistical analysis of the corpus, looking at such things as word frequency and collocational patterns. Comparing surveys in different language variants can help in localization and language harmonization. Apart from the *WageIndicator* questionnaire, the corpus of surveys was translated using the *TRAPD* methodology. *TRAPD* is sometimes called the gold-standard of survey translation.

The authors look at examples from the MCSQ corpus that show how problematic sentences in the past have been translated, in order to inform future translators. Difficulties arise in the use of idioms, particularly where this results in translation variants that differ in scale, leading to non-comparability of survey responses in different languages. For example, the English idiom *A great deal of time* has been variously translated into French as Vraiment beaucoup de temps [Really lots of time], Énormément de temps [Enormous amounts of time], and Une grande partie du temps [A large part of the time].

3. Translation of environmental resources

The second main topic covered by this special issue is the translation of environmental resources. In their paper "Méthodes d'exploitation des corpus pour la traduction de termes complexes" [Methods of corpus exploitation for the translation of complex terms], Melania Cabezas-García and Pilar León-Araúz describe a step-by-step protocol for using corpora to translate multi-word expressions. Translating multi-word terms such as *UV-absorbing aerosol* is one of the main challenges in any translation project. First, the translator must identify and understand them in the source text, but these cannot always be found in terminology banks or dictionaries. The requisite information can often be found in corpora, but many translators are not familiar with them. To remedy this, this paper presents a number of techniques for translating multi-word terms using corpora. The authors illustrate their point by translating English terms into French and Spanish, and make use of the terminological data-base *EcoLexicon* (León-Araúz, Reimerink, *et al.* 2019), which has over 20 million words.

Traditionally, corpora are used as a source to extract term lists, which can be sorted in different ways, such as by frequency, tags, and the words to the right and left of the query. They can also be used to make more fine-grained search requests than one could make on the web. The main part of the paper looks at methods of understanding complex terms using a corpus of the source language; there is also a section on using target language corpora as a translation tool. Dancette (2011) described the notion of *scaffolding* (échafaudage) for the examination of related concepts in the domain by drawing a conceptual system, possibly aided by a taxonomy of terms. This takes the form of a map, where terms are nodes and the relations between them are edges. For example, part of the conceptual system they show is *bottom boundary layer* has_part *sub-layer* generic-of *bottom Elman layer* and *viscous layer*. We can then identify collocations of these words by using the "LogDice" statistic in the *Sketch Engine*[11] concordancer in order to learn about related concepts. The paper also describes searching for elements of complex terms using knowledge patterns similar to Hearst's (1992) templates, such as X is the part of Y; X, composed of Y; and X comprising Y. These allow related terms (Y) to be found for the original term (X, such as *boundary layer*). The frequency of matches in the corpus is a good guide for bracketing a complex term. For example, *bottom boundary* occurs 747 times, *boundary layer*, 24,912 times, and *bottom layer*, 3,569 times. Thus, by bracketing the most frequent word pair, one can parse the complex term as *bottom* [*boundary layer*]. There are also methods for finding multi-word expressions in a target language corpus, such as by searching for equivalent terms in an aligned parallel corpus. For this, the *Sketch Engine* tool has a "Parallel Concordance" option. Another technique is to use the frequency of possible translations of a term in the target language corpus. For example, possible translations of *long-range transboundary air pollution* might be pollution atmosphérique transfrontière, which has 478 hits in Google Scholar, more than any other possible translation, showing that it is the preferred option.

Aurélien Talbot, Camille Biros, and Caroline Rossi's contribution falls within critical corpus-based translation studies (CCTS). It is titled "Pour une traductologie de corpus exploratoire: méthodologie d'analyse d'un corpus de rapports de GIEC et de leurs traductions" [Exploratory corpus-based translation studies: methodology for a corpus analysis of IPCC reports and their translations]. Using the R programming language, the authors perform exploratory statistical analyses of a corpus of English summaries of reports released by the Intergovernmental Panel on Climate Change (IPCC) (*Summaries for Policy Makers*) and their translations into French and Spanish. This is a diachronic study, as the reports span from 1990 to 2014, during which changes in translators' phrase choices have taken place. The corpus contains over 800,000 words in total (with all three languages) and is divided into 5 corresponding assessment reports written in roughly 6-year intervals. The authors state that we are at a turning point, where automation, especially the electronic corpus, has changed translation from a cottage industry into an industrial process.

Trilingual exploratory analysis of IPCC reports was by correspondence analysis, a technique used to represent tabular data in a multi-dimensional mathematical space. Documents and words can be drawn on the same graph, typically in two dimensions, corresponding to the two axes that contribute most to the overall variation in the data. A separate graph was produced for each of the three languages. For each language, on the horizontal axis, the following pattern was seen: the earliest

report, *AR1*, was placed on the left, while the most recent reports, *AR4* and *AR5*, were placed on the right. The other reports appeared in more central positions. The program was set so that words and phrases with loadings > 0.2 (the maximum is 1) were plotted on the same graph as the documents. A number of corresponding terms were placed in equivalent positions for the three languages. For example, the word most closely associated with AR1 was *committee*/comité/**comité**. More acronyms were found near *AR4* and *AR5*, which is typical of a highly specialized discourse. The use of acronyms becomes more prevalent when the field is mature. Differences between languages were also seen. For example, the terms *climate* and *change* do not appear on the English graph, while **cambios** appears on the Spanish graph and both changements and climatique appear on the French one. To explain some of these observations, the authors went on to use a concordance analysis derived from the aligned corpus, which was built from the subcorpora of all three languages. One difference was the word *pathways*, which appeared on the English CA graph and corresponds to **trayectorias** on the Spanish graph; however, there was no equivalent term on the French graph. Annexed to each report was a glossary, where it could be seen that the suggested translations for each of these terms changed from report to report, showing a process of inter-language influence leading to stabilization and clear choices in each language.

4. Classic approaches to CBTS

The remainder of the issue looks at the diverse topics of translation universals (Ebeling), textometrics (Kraif and Roux), and stylometry (McLaughlin).

In "The function of recurrent word-combinations in English translations from three different languages," Signe Oksefjell Ebeling builds on her earlier work, in which she compared fiction translated from Norwegian into English with fiction originally written in English. Here, two other Germanic source languages are used: German and Swedish fiction translated into English. This study contributes to the discussion of translation universals and translation as a third code. Features of translated language that are not shared by the original language are called translationese, the characteristics of which tend to be present irrespective of the source and target languages. This study follows Granger's (2018: 189) rigorous corpus-based methodology of contrastive translation analysis. The experimental results can most likely be explained by source languages "shining through" and the universal tendency for translators to use a smaller and more fixed set of expressions in their translations, examples of simplification and normalization. The source language is now recognized to be an important factor in translation studies after a long time during which people were mainly concerned with translation universals, which are not due to source language interference.

The corpora used were firstly the *English-Norwegian Parallel Corpus* + (ENPC+), an extension of the *English-Norwegian Parallel Corpus*[12] to which a fictional component was added. This was the source of both the texts originally written in English and the Norwegian to English translations. The other two corpora were the *English-Swedish Parallel Corpus* (ESPC)[13] and the *Oslo Multilingual Corpus* (OMC),[14] which includes fiction translated from German into English. The texts in the corpora have all been translated by a range of authors, meaning that individual authorship would

not be a confounding issue when examining the results. A taxonomy was created with 14 classes of phrases, the top levels being evaluative and informational (divided into modalizing and organizational). Examples of criteria for assigning trigrams to a taxonomic type were: a) organizational trigrams contain items that are clearly recognizable as text structuring devices, such as spatial and temporal references; and b) reporting trigrams contain a reporting verb. Trigrams (uninterrupted sequences of three consecutive words) are used to capture phraseological tendencies in the various texts. Using trigrams is a "knowledge-poor" approach that can be easily automated without extensive knowledge of natural languages. Extracted trigrams were both well-dispersed, occurring in at least 25% of the texts, and of a frequency of at least 20 per million words. The trigrams had to meet these criteria in just one of the two texts (English original or translated into English).

This study looks at language pairs that are typologically close and counted the number of trigrams in each taxonomic category for language originals and for translations. For English translations from German, the number of trigrams in eight of the categories in the taxonomy were not significantly different between the German originals and the English translations. Trigrams in two categories were favoured in English originals and trigrams in four categories were more frequent in German to English translation; overall, trigrams in six categories differed significantly. When comparing English originals versus translations from Norwegian, it was again found that 6 out of 14 frequencies were significantly different, and it was also the case for Swedish translated into English. The features that were more common in originals or translations were not always the same: for example, two categories unique to German were existential and report, which were more frequent in English originals. Two categories, comparison and spatial, were significantly different in Norwegian and Swedish and their English translations, but not in English versus German to English translations. This agrees with Ebeling and Ebeling (2018), where they state that "the more frequent use of comparison and spatial 3-grams in English translated from Norwegian is most likely a result of source language shining through." For organisational and temporal trigrams, there may be an explicitation effect, where the translators add references to the organisation of the text and to the times of events. The three source languages thus may thus give a similar "gravitational pull" (Halverson 2017), resulting in explicitations in the case of organisational and temporal trigrams. The overall results give some support for the translation as a third code hypothesis.

While many studies look at the comparison between original texts and their translations, Olivier Kraif and Pascale Roux's paper, "Compairison d'un texte originale et de ses rétrotraductions: que disent les mesures textométriques?" [Comparison of an original text and its back-translation: what do textometric measures tell us?], looks at back-translation, or the translation of a text followed by its retranslation into the original language. They tested three hypotheses: firstly, are texts more greatly transformed when they have been back translated through a language very different from the original? This was tested by back-translation through languages distant from French, such as Japanese, and languages in the same linguistic family, such as Italian or Latin. Secondly, is the distortion due to back-translation greater for poetry than prose? Poetry is often said to be more prone to distortion in translation. Thirdly, can the back-translated corpus contain clues for the interpretation and analysis of the original text?

They used a small sized corpus of texts originally written in French, consisting of a prose essay and a series of four poems. Each of the texts was translated into eight languages, some near to and some distant from French, namely Italian, German, Arabic, Persian, Japanese, Korean, Latin, and Ancient Greek, each by two translators. They were then back-translated by two other translators, who had no knowledge of the texts in their original languages. The original texts and their back-translations were aligned at the sentence, verse, and word level. The first textometric criterion studied by Kraif and Roux was "translational stability." The simplest measure of this was text length or number of tokens. For the essay, this measure was almost always greater for the back-translations than the original, which is consistent with the translation universal of explicitation. The second indicator of translation stability was the use of Dice's similarity coefficient—twice the proportion of tokens shared between the original and the back-translation, divided by the total number of tokens in both texts. Translation stability was also measured based on the number of lemmas that were transformed during back-translation. Kraif and Roux give the analogy of baking a brioche: the raisins, like some lemmas, are not altered in the process, but the rest of the mixture is transformed and recombines like lemmas that change with back-translation. It was found that numerals were the most stable (and pronouns were least stable) for poems, and proper nouns were most stable (and verbs least stable) for the essay.

Mairi McLaughlin's paper, "La traductologie de corpus et la traduction journalistique historique" [Corpus-based translation studies and the translation of historical journalism] looks at both news translation research and the theory of translation universals to determine whether sections of the historical French newspaper, the *Gazette de France*, were originally written in a Germanic language and later translated into French or were French language originals. The section considered more likely to be a translation is "Nouvelles Ordinaires"; experiments were done to see whether this section contained more features associated with translation universals than the main part of the newspaper. The corpus used in this study consisted of all editions of the *Gazette de France* published in January 1632, for a total of five editions each containing 8 pages; it was made up of separate sections called "Gazette" and "Nouvelles ordinaires." The corpus was used to test the hypothesis that the dispatches contained in "Nouvelles ordinaires" had been translated from newspapers in Germanic languages.

The first experiments looked at features of simplification in translation. Language simplification is the tendency for translators to produce lexically and syntactically simpler text in the target language than what was found in the source language. Two features traditionally associated with language simplification are sentence length and lexical diversity, which are both lesser in simpler texts. Mean sentence length was indeed greater for the "Gazette" than "Nouvelles Ordinaires," showing that the latter was constituted of simpler text and therefore more typical of translated text. However, the hypothesis that "Nouvelles Ordinaires" resembled translated text did not hold up to tests measuring lexical density.

Previous studies have shown that explicitation occurs in translated text, when the translator has felt the need to add more information rather than merely make a literal translation, such as when it is necessary to explain cultural words to the intended audience. The characteristics of text said to be indicators of explicitation

were connectors and the use of the passive. There were more conjunctions in "Nouvelles Ordinaires," especially ampersands, confirming the translationese hypothesis. Over-use of the passive was also found in "Nouvelles Ordinaires," the expected finding if explicitation had occurred.

A third translation universal is normalisation, where the translation conforms more to language choices and standards typical of the target language. The language is thus less creative. McLaughlin uses the pronoun on with verbs of reported speech, a standard construction in the early days of French journalism, as an indicator of normalised language. The construction was found to be more frequent in "Nouvelles Ordinaires." A second criterion was called "the tail of the list," or the *hapox legomena*—words which are only seen once in the corpus. These were fewer in "Nouvelles ordinaires," showing that fewer novel lexical choices had been made, and in general, the decision to use more common words had been made. All in all, the three pairs of experiments confirmed the hypothesis that the section called "Nouvelles Ordinaires" featured translations of dispatches originally written in Germanic languages. This work has practical value in the empirical determination of which texts are originals and which are translations, such as in the detection of translation plagiarism.

Last but not least, Mellinger, in his paper "Quantitative questions on big data in translation studies" describes how big data analytic techniques can be productively used in corpus-based translation studies. He shows how modern corpora exhibit the characteristics of big data. The paper distinguishes big data from traditional corpora in terms of the following properties (or V's): volume, variety, velocity, veracity, and value. Three case studies are presented, to show the use of big data methods in CBTS: cross-lingual and multilingual data analysis, sentiment analysis, and visual analysis. The first "V" is volume. Corpora tend to be millions of words in size, but big data sets can be much larger. The *Europarl* corpus is more typical of big data; it was developed to train statistical machine translation systems, which is beyond the scope of general corpora. The second is variety. Corpora have moved beyond consisting of only text to include data such as sign language or images. Corpora may also vary according to how they are annotated or "tagged," such as with part-of-speech tagging or error tagging. A third characteristic of big data is velocity, which refers to the fact that big data sets can be assembled very quickly, such as in the financial sector. This is not generally the case for most corpora, although machine translation can rely on quickly assembled corpora. Veracity and value are more typical of big data sets than general corpora. Value is the ability to improve a product or service, such as providing bespoke translation services. Veracity is the ability to detect biases in big data sets, such as the ability of machine translation systems to provide accurate translations using stored data. Areas of CBTS where big data techniques show promise are cross-lingual and multi-lingual data analysis, sentiment analysis and audio-visual analysis.

5. Conclusion

This special issue illustrates some current approaches to corpus-based translation studies. These include classic CBTS, such as testing translation universal hypotheses using contemporary or historical corpora and textometrics, and some exciting new research approaches, prominently in health and environmental translations. A shared

feature of the papers on health translation is a new focus on the values of translation for readers as individuals with varying reading abilities, for example, the practical understandability of translated health materials or social surveys. This represents a paradigm shift towards a more human-centred approach to translation research, which highlights the diversity and variability among target readers. This stands in contrast with current translation research paradigms, which focus on exploring the underlying, universal patterns in translations. An important contributing factor to the rise of descriptive and, later, corpus translation studies is the fact that they represented a paradigm shift towards a target culture-oriented approach to translation studies, which was highly innovative at the time. It transformed and subverted people's perception of translation as a secondary product of the original work. The search for universal, language-independent patterns of linguistic interventions in translation was to provide evidence of the shared communicative function of translation to interpret and validate the differences between languages and cultural communities. If differences between languages and cultures are not valid and important, translation would become unnecessary. In this sense, descriptive or corpus-based translation studies are not entirely separated from earlier studies that searched for linguistic or functional equivalence between source and target texts. The papers on health translation collected in this special issue represent exciting new research directions prompted by a pressing new social and research topic of our time, widening health inequality. These studies have identified and attempted to fill in a gap in current corpus-based translation studies, that is the suitability of a translated work for readers as valuable individuals within the same target language and culture. This has implicitly challenged the assumption of many studies that the target culture is a uniform whole, overlooking differences at the individual level. The paper by Cabezas-García and León-Araúz, and the one by Talbot, Biros, and Rossi study environmental translations, from the development of multilingual environmental resources to diachronic language use in translated environmental policy materials. Their papers illustrate the impact of multilingual translation data on environmental science and its public and policy communication. As we can see, the papers in this special issue that look at health and environmental translation point to new direction for corpus-based translation studies: health translation paves the way to more human-centred approaches to translation research, and environmental translation illustrates the importance of translation research to other scientific fields and public communication. All papers included in this special issue represent a continuation of corpus-based translation studies, with some exemplifying classic approaches, such as corpus testing for universal translational features, and some heralding a change towards more socially oriented studies that help address pressing social and research questions of our time.

NOTES

1. Bfeh (2021): *Langue facile à lire. Fiche d'information à l'intention de l'administration fédérale*. Version 2.1. Bern: Bureau fédéral de l'égalité pour les personnes handicapées (BFEH). Consulted on 2 February 2022, <https://www.edi.admin.ch/dam/edi/fr/dokumente/gleichstellung/infomaterial/Leichte_Sprache_de_ok.pdf.download.pdf/Langue%20facile%20%C3%A0%20lire.pdf>.
2. Inclusion Europe (Last update: 6 October 2021): Information for all. European standards for making information easy to read and understand. Brussels: European Commission. Consulted on 10 March 2022, <https://www.inclusion-europe.eu/easy-to-read-standards-guidelines/>.

3. Audiau, Aymeric (2009): *L'information pour tous. Règles européennes pour une information facile à lire et à comprendre (FALC)*. Paris: Unipei. Consulted on 21 February 2022, <https://www.unapei.org/wp-content/uploads/2018/11/L%E2%80%99information-pour-tous-Re%CC%80gles-europe%CC%81ennes-pour-une-information-facile-a%CC%80-lire-et-a%CC%80-comprendre.pdf>.
4. Dynamic Quality Framework (DQF) and Multidimensional Quality Metrics (MQM). See Lommel, Arle, Görög, Attila, Melby, Alan, *et al*. (2015): *Harmonised Metric*. Saarbrücken: QT21 Consortium. Consulted on 26 March 2022, <http://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf>.
5. Koehn, Philipp (2005): Europarl: A parallel corpus for statistical machine translation. *In*: Asia-Pacific Association for Machine Translation, dir. *Proceedings of Machine Translation Summit X: Papers*. (MT Summit X: the Tenth Machine Translation Summit, Phuket, 13-15 September 2005). Tokyo: Asia-Pacific Association for Machine Translation, 79-86. Consulted on 19 February 2022, <https://aclanthology.org/2005.mtsummit-papers.11/>.
6. *[MCSQ]: The Multilingual Corpus of Survey Questionnaires* (Last update: 2 August 2021): Consulted on 4 March 2022, <http://easy.mcsq.upf.edu>.
7. *The European Social Survey (ESS)* (Last update: 6 January 2022): Consulted on 27 February 2022, <https://www.europeansocialsurvey.org/>.
8. *European Value Study* (Last update: 30 March 2022): Consulted on 31 March 2022, <https://europeanvaluesstudy.eu>.
9. *SHARE - Survey of Health, Ageing and Retirement in Europe* (Last update: 26 June 2021): Consulted on 3 March 2022, <https://www.share-project.org/>.
10. *Wage Indicator* (Last update: 18 January 2022): Consulted on 26 February 2022, <https://wageindicator.co.uk>.
11. *Sketch Engine* (Last update: 7 March 2022): Consulted on 19 March 2022, <https://www.sketchengine.eu/>.
12. Johansson, Stig, Ebeling, Jarle, and Oksefjell, Signe (1999/2002): *The English-Norwegian Parallel Corpus: Manual*. Oslo: University of Oslo, Department of British and American Studies. Consulted on 7 March 2022, <http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf>.
13. Altenberg, Bengt, Aijmer, Karin, and Svensson, Mikael (2001): *The English-Swedish Parallel Corpus (ESPC)*. University of Lund/University of Göteborg. Consulted on 5 February 2022, <https://www.ipd.gu.se/digitalAssets/1333/1333431_manual_espc.pdf>.
14. Johansson, Stig and Knut Hofland (1994): Towards an English-Norwegian Parallel Corpus. *In*: Udo Fries, Gunnel Tottie, and Peter Schneider, eds. *Creating and Using English Language Corpora*. (ICAME 1993: 14th International Conference on English Language Research on Computerized Corpora, Zurich, 1993). Amsterdam: Rodopi, 25-37.

REFERENCES

Dancette, Jeanne (2011): L'intégration des relations sémantiques dans les dictionnaires spécialisés multilingues: du corpus ciblé à l'organisation des connaissances. *Meta*. 56(2):284-300.

Ebeling, Jarle and Ebeling, Signe Oksefjell (2018): Comparing n-gram-based functional categories in original versus translated texts. *Corpora*. 13(3):347-370.

Felici, Annarita and Griebel, Cornelia (2019): The challenge of multilingual 'plain language' in translation-mediated Swiss administrative communication: a preliminary comparative analysis of insurance leaflets. *Translation Spaces*. 8(1):167-191.

Granger, Sylviane (2018): Tracking the third code. A cross-linguistic corpus-driven approach to metadiscursive markers. *In*: Anna Čermáková and Michaela Mahlberg, eds. *The Corpus Linguistics Discourse. In Honour of Wolfgang Teubert*. Amsterdam/Philadelphia: John Benjamins, 185-204.

Halverson, Sandra (2017): Gravitational pull in translation. Testing a revised model: New Methodological and Theoretical Traditions. *In*: Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, eds. *Empirical Translation Studies. New Methodological and Theoretical Traditions*. Berlin: de Gruyter Mouton, 9-45.

Hareide, Lidun (2019): Comparable Parallel Corpora: A critical review of current practices in corpus-based translation studies. *In*: Irene Doval and M. Teresa Sánchez-Nieto, eds.

Parallel Corpora for Contrastive and Translation Studies. New resources and applications. Amsterdam/Philadelphia: John Benjamins, 19-38.

Harkness, Janet A. (2003): Questionnaire translation. *In*: Janet A. Harkness, Fons J. R. van de Vijver, and Peter Ph. Mohler, eds. *Cross-Cultural Survey Methods*. Hoboken: Wiley and Sons, 35-56.

Hearst, Marti A. (1992): Automatic Acquisition of Hyponyms from Large Text Corpora. *In*: Christian Boitet, ed. *Proceedings of the fifteenth International Conference on Computational Linguistics*. (COLING-92: the 15th International Conference on Computational Linguistics, Nantes, 23-28 August 1992). Nantes: International Committee on Computational Linguistics, 539-545. Consulted on 19 January 2022, <https://aclanthology.org/C92-2082>.

Kuhn, Tobias (2014): A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*. 40(1):121-170.

Laviosa, Sara (1998): Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*. 43(4):557-570.

León-Araúz, Pilar, Reimerink, Arianne, and Faber, Pamela (2019): EcoLexicon and by-products: integrating and reusing terminological resources. *Terminology*. 25(2):222-258.