Meta
Journal des traducteurs
Translators' Journal

META

Quantitative questions on big data in translation studies

Christopher D. Mellinger

Volume 67, numéro 1, avril–mai 2022

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

URI : https://id.erudit.org/iderudit/1092197ar
DOI : https://doi.org/10.7202/1092197ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Mellinger, C. D. (2022). Quantitative questions on big data in translation studies. *Meta*, *67*(1), 217–231. https://doi.org/10.7202/1092197ar

Résumé de l'article

Au fur et à mesure que les études de traduction basées sur les corpus continuent de se développer, les chercheurs ont utilisé des techniques d'analyse de données de disciplines adjacentes telles que la linguistique des corpus pour explorer davantage de questions de recherche. Le domaine a évolué depuis les premières approches basées sur la fréquence jusqu'aux études de traduction basées sur des corpus pour inclure désormais des analyses statistiques plus avancées qui aident à comprendre le réseau complexe de variables qui composent le processus de traduction. Les techniques d'analyse de données massives dérivées de l'analyse des données et des domaines quantitatifs reliés pourraient être appliquées avec succès pour répondre aux questions de recherche des études de traduction et d'interprétation. Pour évaluer leur applicabilité, cet article décrit d'abord ce qui distingue les données massives des corpus généraux dans les études en traduction et en interprétation, en identifiant comment le volume, la variété et la vitesse des données sont des propriétés applicables à prendre en compte dans les études de traduction et d'interprétation basées sur le corpus. L'article présente ensuite trois types de techniques d'analyse des données massives, à savoir l'analyse des données translinguistiques et multilingues, l'analyse des sentiments et l'analyse visuelle. Ces analyses sont présentées conjointement avec les domaines de recherche potentiels qui pourraient bénéficier de ces approches analytiques complémentaires. L'article se termine par une réflexion sur les implications de l'analyse des données massives pour les études de traduction de corpus, tout en décrivant la trajectoire d'une approche plus quantitative, basée sur le corpus, pour les études de traduction.

Quantitative questions on big data in translation studies

CHRISTOPHER D. MELLINGER
The University of North Carolina at Charlotte, Charlotte, USA
cmelling@uncc.edu

RÉSUMÉ

Au fur et à mesure que les études de traduction basées sur les corpus continuent de se développer, les chercheurs ont utilisé des techniques d'analyse de données de disciplines adjacentes telles que la linguistique des corpus pour explorer davantage de questions de recherche. Le domaine a évolué depuis les premières approches basées sur la fréquence jusqu'aux études de traduction basées sur des corpus pour inclure désormais des analyses statistiques plus avancées qui aident à comprendre le réseau complexe de variables qui composent le processus de traduction. Les techniques d'analyse de données massives dérivées de l'analyse des données et des domaines quantitatifs reliés pourraient être appliquées avec succès pour répondre aux questions de recherche des études de traduction et d'interprétation. Pour évaluer leur applicabilité, cet article décrit d'abord ce qui distingue les données massives des corpus généraux dans les études en traduction et en interprétation, en identifiant comment le volume, la variété et la vitesse des données sont des propriétés applicables à prendre en compte dans les études de traduction et d'interprétation basées sur le corpus. L'article présente ensuite trois types de techniques d'analyse des données massives, à savoir l'analyse des données translinguistiques et multilingues, l'analyse des sentiments et l'analyse visuelle. Ces analyses sont présentées conjointement avec les domaines de recherche potentiels qui pourraient bénéficier de ces approches analytiques complémentaires. L'article se termine par une réflexion sur les implications de l'analyse des données massives pour les études de traduction de corpus, tout en décrivant la trajectoire d'une approche plus quantitative, basée sur le corpus, pour les études de traduction.

ABSTRACT

As corpus-based translation studies continues to expand, researchers have employed data analytic techniques from neighbouring disciplines, such as corpus linguistics, to explore a wider variety of research questions. The field has evolved from early frequency-based approaches to corpus-based translation studies to now include more advanced statistical analyses to understand the complex web of variables encapsulated by the translation process. Big data analytic techniques that originated in data analytics and related quantitative fields could be usefully applied to research questions in translation and interpreting studies. To assess their applicability, this article first outlines what distinguishes big data from general corpora in translation and interpreting studies, identifying how data volume, variety, and velocity are applicable properties to be considered in corpus-based translation and interpreting studies research. Then, the article presents three types of big data analysis techniques, namely crosslingual and multilingual data analysis, sentiment analysis, and visual analysis. These analyses are presented in conjunction with potential research areas that would benefit from these complementary analytical approaches. The article concludes with a discussion of the implications of big data analytics in corpus translation studies, while charting the trajectory of a more quantitative, corpus-based approach to translation studies.

RESUMEN

A medida que los estudios de traducción basados en corpus siguen expandiéndose, los investigadores han empleado técnicas de análisis de datos de disciplinas adyacentes como la lingüística de corpus para explorar un mayor número de preguntas de investigación. El campo ha evolucionado desde los primeros enfoques basados en la frecuencia hasta los estudios de traducción basados en corpus para incluir ahora análisis estadísticos más avanzados que permiten comprender la compleja red de variables que integra el proceso de traducción. Las técnicas de análisis de datos masivos que derivaron de la analítica de datos y los campos cuantitativos relacionados podrían aplicarse de forma satisfactoria para dar respuesta a las preguntas de investigación de los estudios de traducción e interpretación. Para evaluar su aplicabilidad, artículo, en primer lugar, esboza la distinción entre los datos masivos y los corpus generales en el contexto de los estudios de traducción e interpretación, centrándose en el volumen, la variedad y la velocidad de los datos como propiedades aplicables a ser consideradas en la investigación de estudios de traducción e interpretación basados en corpus. A continuación, el artículo presenta tres tipos de técnicas de análisis de datos masivos: análisis de datos translingüísticos y multilingües, análisis de sentimientos y análisis visual. Estos análisis se presentan conjuntamente con posibles áreas de investigación que se beneficiarían de estos enfoques analíticos complementarios. El artículo concluye con una reflexión sobre las implicaciones de la analítica de datos masivos para los estudios de traducción de corpus, al mismo tiempo que se perfila la trayectoria de un enfoque más cuantitativo, basado en corpus, para los estudios de traducción.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVE

données massives, recherche quantitative, analyse de données multilingues, analyse des sentiments, analyse audiovisuelle
big data, quantitative research, multilingual data analysis, sentiment analysis, audiovisual analysis
datos masivos, investigación cuantitativa, análisis de datos multilingües, análisis de sentimientos, análisis audiovisual

1. Introduction

Since the inception of corpus-based translation studies, researchers have used quantitative measures to investigate a range of questions about translations as a product. The potential for corpus linguistic research techniques in translation studies was recognised in the early 1990s, and Baker's (1993) essay provided a conceptual roadmap for both theoretical and methodological research[1]. In many cases, research questions centred on so-called universals of translation, comparing translated texts with their source language counterparts (Mauranen and Kujamäki 2004; Malmkjaer 2011). These comparative studies interrogated how these two types of writing – i.e., non-translated and translated texts – differed or corresponded, ranging from lexical and grammatical shifts to more explicit attempts to clarify or disambiguate specific textual features. This line of research continues and is now incorporated into broader considerations of how translators engage with texts, such as cognitive behaviour, stylistics, translation process research, or translator agency (e.g., Hu 2016; Defrancq, Daems, *et al.* 2020).

To conduct these types of studies, researchers regularly rely on frequency data derived from comparable collections of texts. While this type of data (sometimes referred to as count data since, as the name suggests, researchers count the number

of occurrences of a particular string of text or characters) can provide an indication of potential trends in the data, much of the initial work was limited in scope and relied on corpora that, in many instances, were much smaller than those regularly employed in corpus linguistics research. The difference in size is, in many respects, a function of the specialised nature of the texts under consideration; for comparative studies of specific linguistic features, corpora not only need to be comparable but also aligned so that researchers can identify specific occurrences of lexical items to then compare how these were rendered in another language. One challenge of corpus size is striking a balance between being sufficiently small to allow for detailed analysis of specific textual features while being sufficiently large to draw generalisable conclusions beyond the dataset under discussion (for an extended discussion, Malamatidou 2018).

Nevertheless, a reliance on count (frequency) data alone does not establish, in the statistical sense, significant differences or relationships between the various variables of interest. Instead, researchers must rely on more sophisticated data analytic techniques in an effort to systematically investigate specific features of translated texts. For instance, null hypothesis testing is regularly employed in corpus studies to establish probabilistic claims regarding potential differences among variables or to measure the strength of a relationship between specific lexical items (Oakes and Ji 2012; Mellinger and Hanson 2017). Moreover, multivariate analyses have become more widespread in recent years, accounting for the simultaneous co-occurrence of multiple variables in texts (e.g., Kruger 2019). The appearance of larger datasets – e.g., the Europarl parallel corpus or EPTIC[2] – has further required researchers to employ additional statistical techniques, given that these multi-million-word corpora are not feasibly analysed by frequency data alone. The creation of these corpora for translation studies researchers is promising (e.g., Ustaszewski 2019) and has opened greater possibilities to generalise about translated texts in specific domains.

This shift toward big data sources and more sophisticated data analytic techniques is recognised in translation studies now that scholars have greater access to multiple data streams (Mellinger and Hanson 2022). Some scholars have also begun to describe this regular reliance on quantitative approaches as an 'empirical turn' in translation studies (Ji and Oakes 2019). However, for corpus translation studies to be able to fully leverage these new sources, greater consideration of quantitative approaches to data analysis is needed. Consequently, this article argues for the inclusion of big data analytic techniques in corpus translation research in an effort to align corpus-based research questions with appropriate methodological approaches. To do so, the article first distinguishes what constitutes big data and the types of data that may be leveraged in relation to quantitative corpus-based translation studies. Then, specific quantitative methods that are, as of yet, underutilised in the field are discussed in relation to potential areas of interest to propose new avenues of exploration. The article concludes with a discussion of the implications of big data analytics in corpus translation studies, while charting the trajectory of a more quantitative, corpus-based approach to translation studies.

2. Large datasets vs. big data in translation studies

As noted in the introduction, corpus-based translation studies initially found its roots in smaller, specialised corpora to allow specific strings of characters, lexical items, text segments or tokens to be examined in detail. Larger corpora, in contrast, have only recently been compiled, particularly as a result of increased digitisation efforts to make print texts available electronically and of increased computing capacity (Zanettin 2012). However, an important distinction needs to be drawn between large datasets and big data. Whereas these terms both refer to data in quantities that exceed the human capacity to manually record or analyse, big data is generally described by researchers and data analysts in terms of properties referred to as the 3 Vs: volume, variety, and velocity (Laney 2001). Additional properties have since been added to the initial three, namely veracity and value, representing an evolving understanding of how big data and its analysis can help better understand vast amounts of data beyond what is possible through human review alone (Jin, Wah, *et al.* 2015). Considered together, these properties provide a useful framework to discuss big data in translation studies.

The first property, *volume*, is most readily understood as the quantity of data with which researchers are working. In corpus-based translation studies, volume is most often described as the number of words or segments included in a corpus. Larger comparable or aligned corpora provide millions of words that researchers can query to investigate specific research questions related to the act of translation. The types of questions that can be asked are necessarily constrained by corpus composition; for instance, the large multilingual, aligned corpus Europarl is better suited for questions related to institutional translation than to literary translation (Islam and Mehler 2012). In contrast, the Translational English Corpus (TEC)[3] comprises four main text types – i.e., fiction, biography, news and inflight magazines – and is more suited for questions related to these subcorpora (Baker 1993; 1995). These corpora are sufficiently large to allow researchers to generalise beyond the texts included in the corpora. In fact, Europarl was initially developed by Koehn (2005) in an effort to develop and train statistical machine translation systems.

Volume, though, is not the only property of interest to big data analysts; *variety* is an important property that describes the range of potential data sources and data types that are available to researchers. Within the realm of corpus-based translation studies, this property is perhaps viewed most commonly with respect to text types or genres. Yet data variety ought to be considered more broadly in its applicability to translation and interpreting studies researchers. For instance, corpus-based interpreting studies scholars have begun to investigate how signed language interpreting corpora may be compiled and analysed, eschewing the text-only approach to corpus-based approaches that have, to date, prevailed in the field in order to incorporate data that are not solely written texts (Wehrmeyer 2019).

Corpus data may also exhibit variety with respect to structure. In some cases, data are structured by means of tokenization, segmentation or textual alignment. The level to which texts are structured is also variable, insofar as some compilers may want to have a highly segmented corpus to allow for detailed micro-analyses while others may choose to align corpora using larger sections of text in the interest of time, speed, and resource management. In other cases, relationships are explicitly established among

data points through the use of meta-data or associated tags. The use of tags in translation corpus research has been done in a variety of ways, including parts of speech or error tagging as well as semantic annotation (Zanettin 2000; 2013). While some types of tagging can be automated, other types must be manually inserted by researchers for their specific purposes (for a more detailed overview, see Zanettin 2000). Researchers can also use tags to restructure or 'clean' datasets to address specific research questions. For example, Ustaszewski (2019) used existing tags contained in the Europarl corpus, such as the name of the speaker, the language of the speaker and the language of the text to identify translation directionality. Since the Europarl corpus was already structured, but failed to account for directionality, Ustaszewski (2019) found it necessary to clean the data or find ways to pre-process the corpus in a manner that allowed for a more nuanced analysis with respect to specific variables. Since the researcher was interested in directionality, restructuring the corpus in this way provides a means for future researchers to investigate specific translation features or shifts while still controlling for directionality. The example of Ustaszewski (2019) is but one of a growing number that rely on tagged corpora, and there is still considerable space to query structured and linked data. Bowker and Delsey (2016) recognise the potential for linked data to advance research agendas within both translation studies and information science and advocate for collaboration in these areas.

In many cases, however, textual data are not structured, so analytic techniques are needed that specifically address unstructured data sources. As noted above, researchers can impose a structure by either manually or automatically coding and tagging a text. While the specifics of these techniques lie outside the scope of this article, researchers must be mindful that this initial pre-processing of data needs to be done within a theoretical framework that will allow for the subsequent analyses needed to answer the posed research questions.

The third property of big data, *velocity*, refers to the speed with which data are produced and analysed; big data research is characterised by situations in which data are generated very quickly and/or in large quantities. For example, consider financial markets or online vendors and the speed at which transactions are made along with all of the associated information for each transaction. With millions of daily transactions, one can quickly imagine the overwhelming size of the data streams and the speed at which they are created. Corpus-based translation studies has yet to contend with data streams on this scale; however, machine translation and speech-to-speech translation research increasingly relies on large data streams being processed in real-time, which may require big data analytic approaches that are used in other fields (Kowalski 2016; Nguyen, Stüker, *et al.* 2020). In a similar vein, researchers working with corpus-building software that employ webscraping, web crawling, and bootstrapping technologies to quickly compile large monolingual or multilingual corpora are likely to encounter similarly large volumes of content (see, for instance, Toral, Esplá-Gomis, *et al.* 2016). Moreover, big data analytic techniques may allow for this type of work to be done more readily within the translation studies community, particularly for those working with social media data streams that are generated quickly and in large quantities.

The final two properties that have been described by big data scientists, namely *veracity* and *value*, are more focused on the application of big data analysis. Veracity refers to the ability of big data algorithms to identify potential biases in data and

make predictions that are likely to be true on the basis of data input. To use machine translation again as an example, the ability for machine translation (MT) engines to generate sufficiently accurate predictions on the basis of stored data is one way to view veracity. Value is related to the use of big data to improve a product or service. In the language service industry, for instance, this focus might be on tailoring translation services for specific clients (Koskinen 2020). In a similar vein, big data may add value for researchers working in applied areas, leveraging analytical techniques that can automatically process crosslingual or multilingual resources to describe or predict translation and interpreting. This type of real-time feedback loop of data collection, processing, analysis, and re-integration of data would represent a shift in how corpus-based research is commonly conducted in translation studies, ultimately opening new avenues for research. To do so, however, requires adapting big data approaches to research to the specific challenges and needs of the translation and interpreting studies communities. In what follows, these big data approaches are first outlined, followed then by three areas of translation studies that are particularly well suited to big data analytics. These research areas – namely cross-/multilingual data analysis, sentiment analysis, and visual analysis – represent several areas of translation and interpreting (T&I) investigation, and for which big data techniques are appropriate. These complementary analytical approaches are well suited for research questions in those areas and are prime for consideration.

3. Adapting big data approaches to translation studies

The idea to adapt big data approaches to a new area of research is not new. Data science, as a discipline, does not have datasets of its own, and instead encompasses a series of analytical techniques and research approaches that can be applied across many disciplines (Slota, Hoffman, *et al.* 2020). For instance, DiMaggio (2015) describes how computational text analysis may be well suited for the social sciences, while Chen and Wojcik (2016) describe how big data approaches may be adapted for the psychological sciences. Others, such as Mahmoodi, Leckelt, *et al.* (2017) highlight the value of big data analytic techniques when integrated into the social and behavioral sciences, while recognising the concessions that researchers must make when adopting a different research paradigm. In fact, a full special issue of *Psychological Methods*, edited by Harlow and Oswald (2016), specifically addresses many of the aforementioned topics. Slota, Hoffman, *et al.* (2020) identifies prospecting for data sets as a main challenge – i.e., the identification of potential data sources that can be ordered, structured, or understood. Additionally, the ability to access these data sources, even by researchers within the specific research community or population of interest, can be mired by challenges related to confidentiality, privacy, ethics, and security (Richards and King 2014; Mellinger 2020). This type of work ultimately must be embedded within a discipline's epistemological approach to research and in line with the theoretical frameworks and research questions driving these studies. Yet the possibility of data scientists engaging with T&I scholars represents a potential area of growth, particularly within the corpus-based translation and interpreting studies communities.

While the analytical techniques vary, many researchers distil these into four main types of data analysis: descriptive; diagnostic or causal; predictive; and prescriptive (for a brief overview, see Holmes 2017). As these terms suggest, big data analysis

can seek to describe data, establish causal relationships between variables, predict potential outcomes based on previous behaviour or data, and prescribe an optimal course of action. In the context of corpus-based translation studies, much of the work in the field has been descriptive, focusing on the characteristics of texts that have been translated or interpreted. Some researchers have also used these descriptions to try to establish potential relationships between textual features and the target language rendition or predict behaviour on the basis of previous work. These categories, while useful, are somewhat nebulous and might be considered on a cline rather than being mutually exclusive. Aggarwal (2015: xxiii) conceptualises big data approaches somewhat differently, and instead focuses on four 'super problems' that big data analysis can address: clustering, classification, association pattern mining, and outlier analysis. These categories focus more on the analytical techniques and lend themselves a bit more to the discussion in subsequent sections of this article.

Another aspect of data analysis that does not neatly fit into these categorisations is data visualisation. Techniques to represent big data sets visually can be used descriptively to provide potential clues about relationships among various data points. For instance, a graphical representation of data may reveal potential clusters or outliers that require closer analysis (Mellinger and Hanson 2017). Likewise, visual mappings of networks may uncover relationships that may not be easily identified or described in text alone, and instead provide another means by which to understand relationships among various data points (e.g., McCarty, Molina, *et al.* 2007). Visualisation techniques can also be used to present research findings, synthesising large amounts of data into accessible visual infographics or creating interactive dashboards.

3.1. Crosslingual and multilingual data analysis

As noted above, much of the corpus-based research in translation and interpreting studies has been descriptive, allowing researchers to identify and compare textual features of texts across multiple languages. The intersection of big data analytical approaches and these corpus-based studies is succinctly described by Steiner (2017) in his discussion of cross-fertilisation of computational linguistics and translation studies, akin to some of the discussion here related to the inclusion of big data approaches to translation studies research. Steiner contends that theoretical and methodological divides could be bridged using contrastive corpus analyses to examine explicitness and explicitation, in line with notions of translation universals and some of the initial forays of translation studies scholars into corpus studies. Steiner also argues that textual cohesion is another area that merits investigation to draw on similar corpus architectures as those required by contrastive studies to better understand causal relationships among variables. These studies move beyond descriptive approaches to corpus research that rely on co-variation hypotheses and instead seek to understand the influence of one text on another. The final area he describes is an effort to bridge product- and process-oriented studies, relying on triangulated data sources to understand how these variables are interrelated.

Steiner's (2017) discussion focuses largely on text-based analyses and highlights the need for theoretically-driven work to conduct this type of crosslinguistic analysis. For instance, from a translation studies perspective, Kruger (2019) uses this type

of comparative approach to examine two hypotheses, namely the risk-aversion hypothesis and cognitive complexity hypothesis, to understand translator behaviour and the possibility that translators opt for what might be considered clearer writing in English by using *that* as a complementizer. By linking specific textual features to a theoretical framework, Kruger is able to examine explicitness as a potentially subconscious approach to translation (see Olohan and Baker 2000). In a similar approach, Patton and Can (2012) attempt to identify invariant characteristics by looking at translations of Jame Joyce's *Dubliners*. This type of analysis examines style in translation and how a source text may influence its target text rendition. These theoretically-grounded studies illustrate how larger questions related to cognitive processes or stylistics can be operationalised within a corpus study, thereby illustrating how difficult-to-define constructs can be instantiated within a text.

Big data analytic techniques allow for similar types of studies, for instance in relation to ideas of authorship, knowledge flows in translation, network analyses, and plagiarism. These concepts are linked, insofar as the creation of a text in one language and its circulation in others can be dependent on networks, translation, and the unfortunate reality of uncredited appropriation. When conducted manually, this type of work can be quite laborious and time-consuming and often must be based on specific texts for close readings and comparisons. However, researchers have begun to use big data techniques to automate plagiarism detection. For instance, Ezzikouri, Oukessou, *et al.* (2018) describe how large-scale text comparison based on semantic similarity can be quite challenging in light of the sheer volume of texts available and the propensity for plagiarists to modify or adapt automatically translated texts to avoid detection. To mitigate for the challenges inherent in manual comparison, Ezzikouri, Oukessou, *et al.* (2018) employ big data to automate this detection process. The methods used for this type of analysis vary (see, for instance, Barrón-Cedeño, Gupta, *et al.* 2013 for an overview), yet the ability to systematically process texts may detect instances of semantic similarity or plagiarism that had previously gone unrecorded. This type of research intersects with questions surrounding copyright as well, not only with respect to written texts, but also audiovisual material that may be pirated or adapted without permission of the copyright holder (Gray and Suzor 2020). Moreover, translation studies researchers who are interested in this area of work might use these big data techniques to initially identify candidates for more detailed, manual analyses or comparisons.

Two additional big data analytical techniques that may enhance crosslingual research are clustering and outlier analysis. These techniques, as their names suggest, allow researchers to identify specific concepts or terms that appear in close proximity and those that appear to be aberrant or out of place. In corpus-based studies, the ability to identify clusters has been explored at length (Moisl 2015) and seeks to understand how 'close' items are. Defining closeness is not simply a matter of position within a text, but rather often relies on different mathematical conceptualisations of distance (e.g., Euclidean distance) to measure how far specific items may be. These approaches may be well suited for understanding similarities across translations that move beyond lexical comparisons to instead account for semantic webs or networks. The other analytic technique, outlier analysis, provides a window into lexical items, character strings or text segments that may be out of place in a source text or a translation. Corpus studies have looked at this from the perspective of singularly-

occurring items (i.e., hapax legomena) as well as using more sophisticated models (for a review of several models, see, for instance, Kannan, Woo, *et al.* 2017). These models might be useful to determine stylistic differences across multiple translators, to separate texts or articles that were translated by specific individuals, or to identify aberrant behaviour for closer inspection. Again, these big data approaches to text-based analysis may provide a starting point for researchers to identify specific instances that merit additional research while automating some of the processing algorithms.

3.2. Sentiment analysis

A second area of translation and interpreting studies research that is well suited to big data analytic techniques is sentiment analysis. This type of research focuses on understanding what emotions or feelings are present or associated with specific stimuli. Some researchers working with sentiment analysis have identified specific emotional valences associated with lexical items (e.g., Stadthagen-Gonzalez, Imbault, *et al.* 2017), while others have focused on identifying sentiment and opinions in social media (Pak and Paroubek 2010) or more generally in text (Chatterjee, Gupta, *et al.* 2018). In the case of corpus-based translation and interpreting studies, researchers have focused on how emotions are rendered in another language or altered during the translation process (e.g., Ji and Oakes 2012) as well as how they appear in metaphors to understand how emotions and imagery are construed in translation (e.g., Lewandowska-Tomaszczyk 2012).

A potential site at which sentiment analysis and translation studies intersect is social media and its use as a corpus. For instance, Desjardins (2017) discusses emotion in social media in relation to the use of emojis, which, in some respects, can be considered a form of tag or lexical item that provides structure to data. Zappavigna (2018) makes this more explicit in her analysis of hashtags as a metadiscourse, with both of these authors describing how these discoursal strategies are used in language, be it in original writing or in translation. Still others have considered emoticons to be nonverbal cues of communication, seeking to understand how these function in multiple languages (e.g., Park, Baek, *et al.* 2014). With the growing interest in crowdsourced and non-professional translation as well as studies on translator or interpreter attitudes and beliefs, social media feeds on Twitter, Facebook, and LinkedIn are prime data sources that have yet to be fully explored.

As in the previously-described crosslingual data analysis section, this type of work is time-intensive and is likely to be incomplete without the use of specialised tools or software. Consequently, big data analyses that automate this work are of considerable interest. Recent research by Salameh, Mohammad, *et al.* (2015) demonstrates the potential for this type of work by using sentiment prediction to analyse social media posts written originally in Arabic and then translated into English using both human and machine translation. Their findings are suggestive that human translation is more susceptible to shifts in sentiment with respect to social media posts than machine translation systems. This finding may not be particularly surprising, insofar as many machine translation systems and sentiment analysis algorithms are built on assumptions derived from textual input at a superficial level rather than on a deeper semasiological understanding of the texts. Nevertheless, the ability to assess translations with respect to sentiment may allow researchers to complement phenomenological readings

of a text for emotional valence with automated, systematised analyses. In doing so, larger bodies of text can be examined with translation functioning as the primary variable to understand, at least at the textual level, how sentiment is altered via translation (e.g., Brooke, Tofiloski, *et al.* 2009; Mohammad, Salameh, *et al.* 2016).

3.3. Audiovisual analysis

The two previous sections have focused primarily on text-based corpus studies that ultimately rely on digital texts for analysis; however, corpora comprising images and visual media represent a significant lacuna in translation and interpreting studies research. This gap is somewhat surprising, given the recent interest in imagology (e.g., van Doorslaer, Flynn, *et al.* 2016), visual representations of interpreters (e.g., Fernández-Ocampo and Wolf 2014), and intersemiotic modalities of translation (e.g., Desjardins 2008; Pereira 2008). Several efforts to create corpora for signed language interpreting research (Wehrmeyer 2019) and audiovisual translation (Baños, Bruti, *et al.* 2013) have resulted in corpora that rely on textual analysis and tagging; however, big data analytic techniques may augment the methodological tools available to T&I researchers working with these multimodal corpora.

For instance, clustering analyses are useful to group visual data based on similarities. Studies that have employed this technique are scarce in translation and interpreting studies, but there are a number of potential avenues worth exploring. For instance, researchers working on visual representations of translators and interpreters may try to categorise images into thematic categories. This inductive, manual approach to data analysis is, by its very nature, iterative, and may lead to potential inconsistencies in categorisation. In many contexts, the use of multiple coders can mitigate for this challenge, but automated visual analyses may be another option (Zhang, Stoffel, *et al.* 2012). These categories can be refined and adjusted by researchers, but this initial pass may yield significant time savings and allow researchers to pinpoint specific outliers or categories that occur most often in data.

More deductive approaches to visual analysis are also possible using these analytic techniques. For instance, visual classification allows researchers to train a system using specific images to impose thematic categories or allow researchers to model what might be considered prototypical images of specific items (Zhang, Stoffel, *et al.* 2012). In the case of translation and interpreting studies, this approach might take the form of identifying spatial distance between interpreters and interlocutors, understanding the body language of people in the images, or detecting a specific object in the image. Researchers have also used visual data to recognise emotion in facial expressions (e.g., Ruiz-Garcia, Elshaw, *et al.* 2016), which may provide additional data streams to triangulate interpreter performance with the emotional valence of the situations in which they work. The same holds true for audio content analysis, which provides for different types of sound to be classified on the basis of specific parameters (e.g., Zhang and Kuo 2001). This type of analysis may help operationalise what constitutes a specific speech pattern or timbre, music or sounds in a film or television program, or background noise or distractors underneath spoken language. These types of analyses may provide groupings that facilitate analysis vis-à-vis other variables, such as interpreter performance, or analyses that account for multimodal translation, such as in the case of subtitling and surtitling. The multimodal nature of

these studies ultimately requires simultaneous consideration of all of these variables, and big data analytic techniques provide a means to automate some of this process.

4. Conclusion

The three previous sections, namely crosslingual and multilingual analysis, sentiment analysis, and audiovisual analysis, are overviews of potential areas in which translation and interpreting scholars may benefit from big data analytic techniques. In many instances, researchers have already established theoretical frameworks and models within which these topics can be further explored, allowing for an epistemological alignment with the theoretical and methodological approaches that big data research can encompass. The benefits of this type of research have already been seen by scholars working in machine translation and natural language processing, leveraging big data frameworks to develop, statistical and neural machine translation systems that eclipse translation outputs from early rule-based translation systems (Koehn 2020). Moreover, the ability to integrate these systems with speech algorithms has facilitated additional avenues of study, such as speech-to-speech translation. In a similar vein, translation process researchers have identified the utility of big data analytic approaches to analyse behavioural data derived from translators and interpreters (see, for instance, Carl, Bangalore, *et al.* 2016). In addition, the regular engagement of translators with cloud-based big data systems has shifted the ways we prepare students to work in professional contexts (Wang 2019), requiring a macrolevel view of how translation data are processed, analysed, and used.

The described big data techniques, however, are by no means infallible, nor is the list presented here exhaustive. Aggarwal and Zhai (2012) present an overview of many of the types of analyses that can be conducted in this rapidly evolving area of research. Moreover, researchers working with these approaches recognise the need for regular intervention and refinement by researchers to understand the complex variables that are in play. Nevertheless, this oversight should not obviate the potential for big data approaches in any of these areas to provide starting points for researchers as they seek to understand questions that lie at the intersection of textual, visual, and aural data streams. Corpus-based translation and interpreting studies are now well positioned to leverage these types of analytical techniques and their use will ultimately open the possibilities of what research questions can be posed and interrogated while questioning the extent to which results can be generalised to translation and interpreting.

NOTES

1. In a similar vein, corpus-based interpreting studies has developed almost in parallel, initially described by Shlesinger (1998). In this article, the term *corpus-based translation studies* will be used to encompass both translation and interpreting.
2. For a more detailed description of the development and compilation of the European parallel corpus, see Koehn (2005). For the European Parliament Translation and Interpreting Corpus (EPTIC), see the work of Bernardini (2016).
3. The Translational English Corpus (TEC) was originally established by Mona Baker. Additional information about the corpus is available online at http://genealogiesofknowledge.net/translational-english-corpus-tec/. Recent functionality improvements for data analysis are reported in Luz and Sheehan (2020).

REFERENCES

Aggarwal, Charu C. (2015): *Data Mining: The Textbook*. Cham, Switzerland: Springer.

Aggarwal, Charu C. and Zhai, ChengXiang, eds. (2012): *Mining Text Data*. Singapore: Springer.

Baker, Mona (1993): Corpus linguistics and translation studies. *In*: Mona Baker, Gill Francis, and Elena Tognini-Bonelli, eds. *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 233-250.

Baker, Mona (1995): Corpora in translation studies: An overview and some suggestions for future research. *Target*. 7(2):223-243.

Baños, Rocío, Bruti, Silvia, and Zanotti, Serenella (2013): Corpus linguistics and audiovisual translation: In search of an integrated approach. *Perspectives: Studies in Translation Theory and Practice*. 21(4):483-490.

Barrón-Cedeño, Alberto, Gupta, Parth, and Rosso, Paolo (2013): Methods for cross-language plagiarism detection. *Knowledge-Based Systems*. 50:211-217.

Bernardini, Silvia (2016): Intermodal corpora: A novel resource for descriptive and applied translation studies. *In*: Gloria Corpas Pastor and Miriam Seghiri, eds. *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*. Frankfurt: Peter Lang, 129-148.

Bowker, Lynne and Delsey, Tom (2016): Information science, terminology and translation studies: Adaptation, collaboration, integration. *In*: Yves Gambier and Luc van Doorslaer, eds. *Border Crossings: Translation Studies and Other Disciplines*. Amsterdam: John Benjamins, 73-96.

Brooke, Julian, Tofiloski, Milan, and Taboada, Maite (2009): Cross-linguistic sentiment analysis: From English to Spanish. *International Conference RANLP 2009*. 50-54.

Carl, Michael, Bangalore, Srinivas, and Schaeffer, Moritz (2016). Computational linguistics and translation studies. *In*: Yves Gambier and Luc van Doorslaer, eds. *Border Crossings: Translation Studies and Other Disciplines*. Amsterdam: John Benjamins, 225-244.

Chatterjee, Ankush, Gupta, Umang, Chinnakotla, Manoj Kumar, *et al*. (2018): Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*. 93:309-317.

Chen, Eric Evan and Wojcik, Sean P. (2016): A practical guide to big data research in psychology. *Psychological Methods*. 21(4):458–474.

Defrancq, Bart, Daems, Joke, and Vandevoorde, Lore, eds. (2020): *New Empirical Perspectives on Translation and Interpreting*. New York: Routledge.

Desjardins, Renée (2008): Intersemiotic translation and cultural representation within the space of the multi-modal text. *TranscUlturAl*. 1(1):48-58.

Desjardins, Renée (2017): *Translation and Social Media: In Theory, In Training and In Professional Practice*. London: Palgrave.

DiMaggio, Paul (2015): Adapting computational text analysis to social science (and vice versa). *Big Data & Society*. 2(2):1-5.

Ezzikouri, Hanane, Oukessou, Mohamed, Madani, Youness, *et al*. (2018): Fuzzy cross language plagiarism detection (Arabic-English) using WordNet in a big data environment. *ICCBDC'18: Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing*. 22-27.

Fernández-Ocampo, Anxo and Wolf, Michaela, eds. (2014): *Framing the Interpreter: Towards a Visual Perspective*. New York: Routledge.

Gray, Joanne E. and Suzor, Nicolas P. (2020): Playing with machines: Using machine learning to understand automated copyright enforcement at scale. *Big Data & Society*. 7(1):1-13.

Harlow, Lisa L. and Oswald, Frederick L. (2016): Big data in psychology: introduction to the special issue. *Psychological Methods*. 21(4):447-457.

Holmes, Dawn E. (2017): *Big Data: A Very Short Introduction*. Oxford: Oxford University Press.

Hu, Kaibao (2016): *Introducing Corpus-Based Translation Studies*. London: Springer.

Islam, Zahurul and Mehler, Alexander (2012): Customization of the Europarl corpus for translation studies. *In*: Nicoletta Calzolari, *et al*., eds. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul: ELRA, 2505-2510.

Ji, Meng and Oakes, Michael J. (2012): A corpus study of early English translations of Cao Xueqin's *Hongloumeng*. *In*: Michael J. Oakes and Meng Ji, eds. *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, 177-208.

Ji, Meng and Oakes, Michael J. (2019): Challenges and opportunities of empirical translation studies. *In*: Meng Ji and Michael Oakes, eds. *Advances in Empirical Translation Studies*. Cambridge: Cambridge University Press, 252-264.

Jin, Xiaolong, Wah, Benjamin W., Cheng, Xueqi, *et al*. (2015): Significance and challenges of big data research. *Big Data Research*. 2(2):59-64.

Kannan, Ramakrishnan, Woo, Hyenkyn, Aggarwal, Charu C., *et al*. (2017): Outlier detection for text data. *Proceedings of the 2017 Siam International Conference on Data Mining*. 489-497.

Koehn, Philipp (2005): Europarl: A parallel corpus for statistical machine translation. *Conference Proceedings: The Tenth Machine Translation Summit*. Phuket, Thailand: MT Summit, 79-86.

Koehn, Philipp (2020): *Neural Machine Translation*. New York: Cambridge University Press.

Koskinen, Kaisa (2020): Tailoring translation services for clients and users. *In*: Erik Angelone, Maureen Ehrensberger-Dow, and Gary Massey, eds. *The Bloomsbury Companion to Language Industry Studies*. London: Bloomsbury, 139-152.

Kowalski, Maciej (2016): Learning curve with machine translation based on parallel, bilingual corpora. *In*: Dominik Ryzko, *et al*., eds. *Machine Intelligence and Big Data in Industry*. Cham, Switzerland: Springer, 11-22.

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.

Laney, Doug (2001): 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*. 6:70–73.

Lewandowska-Tomaszczyk, Barbara (2012): Explicit and tacit: An interplay of the quantitative and qualitative approaches to translation. *In*: Michael J. Oakes and Meng Ji, eds. *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, 1-34.

Luz, Saturino and Sheehan, Shane (2020): Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge. *Palgrave Communications*. 6: Article 49.

Mahmoodi, Jasmin, Leckelt, Marius, van Zalk, M.W.H., *et al*. (2017): Big Data approaches in social and behavioral science: four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*. 18:57-62.

Malamatidou, Sofia (2018): *Corpus Triangulation: Combining Data and Methods in Corpus-based Translation Studies*. New York: Routledge.

Malmkjaer, Kirsten (2011): Translation universals. *In*: Kirsten Malmkjaer and Kevin Windle, eds. *The Oxford Handbook of Translation Studies*, Oxford: Oxford University Press, 83-94.

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

McCarty, Christopher, Molina, José Luis, Aguilar, Claudia, *et al*. (2007): A comparison of social network mapping and personal network visualization. *Field Methods*. 19(2):145-162.

Mellinger, Christopher D. (2020): Core research questions and methods. *In*: Erik Angelone, Maureen Ehrensberger-Dow, and Gary Massey, eds. *The Bloomsbury Companion to Language Industry Studies*. London: Bloomsbury, 15-35.

Mellinger, Christopher D. and Hanson, Thomas A. (2017): *Quantitative Research Methods in Translation and Interpreting Studies*. New York: Routledge.

Mellinger, Christopher D. and Hanson, Thomas A. (2022): Research data. *In*: Federico Zanettin and Christopher Rundle, eds. *Routledge Handbook of Translation and Methodology*. New York: Routledge, 307-323.

Mohammad, Saif M., Salameh, Mohammad, and Kiritchenko, Svetlana (2016): How translation alters sentiment. *Journal of Artificial Intelligence Research*. 55:95-130.

Moisl, Hermann (2015): *Cluster Analysis for Corpus Linguistics*. Berlin: Walter de Gruyter.

Nguyen, Thai-Son, Stüker, Sebastian, Niehues, Jan, *et al.* (2020): Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain*. 7689-7693.

Oakes, Michael J. and Ji, Meng, eds. (2012): *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins.

Olohan, Maeve and Baker, Mona (2000): Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*. 1(2):141-158.

Pak, Alexander and Paroubek, Patrick (2010): Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: ELRA, 1320-1326.

Park, Jaram, Baek, Young Min, and Cha, Meeyoung (2014): Cross-cultural comparison of nonverbal cues in emoticons on Twitter: Evidence from big data analysis. *Journal of Communication*. 64:333-354.

Patton, Jon M. and Can, Fazli (2012): Determining translation invariant characteristics of James Joyce's *Dubliners*. *In*: Michael Oakes and Meng Ji, eds. *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*. Amsterdam: John Benjamins, 209-229.

Pereira, Nilce M. (2008): Book illustration as (intersemiotic) translation: Pictures translating words. *Meta*. 53(1):104-119.

Richards, Neil M. and King, Jonathan H. (2014): Big data ethics. *Wake Forest Law Review*. 49(1):393-432.

Ruiz-Garcia, Ariel, Elshaw, Mark, Altahhan, Abudulrahman, *et al.* (2016): Deep learning for emotion recognition in faces. *In*: Alessandro E.P. Villa, Paolo Masulli, and Antonio Javier Pons Rivero, eds. *Artificial Neural Networks and Machine Learning – ICANN 2016*, Part II. Cham, Switzerland: Springer, 38-46.

Salameh, Mohammad, Mohammad, Saif M. and Kiritchenko, Svetlana (2015): Sentiment after translation: A case-study on Arabic social media posts. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. Denver, CO: ACL, 767-777.

Shlesinger, Miriam (1998): Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta*. 43(4):1-8.

Slota, Stephen C., Hoffman, Andrew S., Ribes, David, *et al.* (2020): Prospecting (in) the data sciences. *Big Data & Society*. 7(1):1-12.

Stadthagen-Gonzalez, Hans, Imbault, Constance, Pérez Sánchez, Miguel A., *et al.* 2017. Norms and valence and arousal for 14,031 Spanish words. *Behavior Research Methods*. 49:111-123.

Steiner, Erich (2017): Methodological cross-fertilization: Empirical methodologies in (computational) linguistics and translation studies. *In*: Oliver Czulo and Silvia Hansen-Schirra, eds. *Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation, TC II*. Berlin: Language Science Press, 65-90.

Toral, Antonio, Esplá-Gomis, Miquel, Klubička, Filip, *et al.* (2016): Crawl and crowd to bring machine translation to under-resourced languages. *Language Resources & Evaluation*. 51:1019-1051.

Ustaszewski, Michael (2019): Optimising the Europarl corpus for translation studies with the EuroparlExtract toolkit. *Perspectives: Studies in Translation Theory and Practice*. 27(1):107-123.

Van Doorslaer, Luc, Flynn, Peter, and Leerssen, Joep, eds. (2016): *Interconnecting Translation Studies and Imagology*. Amsterdam: John Benjamins.

Wang, Huashu (2019): The development of translation technology in the era of big data. *In*: Feng Yue, *et al.*, eds. *Restructuring Translation Education*. Singapore: Springer, 13-26.

Wehrmeyer, Ella (2019): A corpus for signed language interpreting research. *Interpreting*. 21(1):62-90.

Zanettin, Federico (2000): Parallel corpora in translation studies: Issues in corpus design and analysis. *In*: Maeve Olohan, ed. *Intercultural Faultlines: Research Models in Translation Studies, Vol. 1*. London: Routledge, 105-118.

Zanettin, Federico (2012): *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. New York: Routledge.

Zanettin, Federico (2013): Corpus methods for descriptive translation studies. *Procedia: Social and Behavioral Sciences*. 95:20-32.

Zappavigna, Michele (2018): *Searchable Talk: Hashtags and Social Media Discourse*. London: Bloomsbury.

Zhang, Leishi, Stoffel, Andreas, Behrisch, Michael, *et al*. (2012): Visual analytics for the big data era – A comparative review of state-of-the-art commercial systems. *IEEE Symposium on Visual Analytics Science and Technology*. Seattle, WA: IEEE, 173-182.

Zhang, Tong and Kuo, C.-C. Jay (2001): Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*. 9(4):441-457.