THE PAST, PRESENT, AND FUTURE OF CORPUS AND TRANSLATION STUDIES

Niall Curry and Tony McEnery

0.1 The Development of Translation Studies

At its core, the field of translation studies is dedicated to the exploration of translation phenomena, encompassing the process of transferring meaning from one language (the source) to one or more other languages (the target). Such phenomena have long been of interest to scholars, with, for example, the ancient Roman philosopher Cicero and poet Horace being credited with proposing the initial critiques of word-for-word and sense-for-sense translation approaches–methods later interrogated and applied by St. Jerome in his work on the *Biblia Vulgata*, a Latin translation of the Bible. These contributions notwithstanding, the field, as we understand it today, began to take shape in earnest in the second half of the twentieth century, with the seminal work of scholars such as Nida (1964), developing their 'science of translation' and Holmes (1972), proposing the term 'translation studies' to describe this new branch of knowledge. Nowadays, the relevance of translation studies extends beyond literary and linguistic borders and research in translation studies has become crucial for understanding cross-cultural communication, international relations, and the global dissemination of scientific knowledge.

As the field of translation studies developed, wider movements in linguistics served to shift the focus from 'langue' to 'parole', in the Saussurean sense, where translation practices required not only an understanding of the mechanics of the source and target languages, but also a contextualised positioning of the text being translated and the context or contexts into which it was being translated. This dual focus led to the emergence of synergetic relationships with a number of parallel fields in linguistics, most prominently linguistic typology and contrastive linguistics.

As linguistic typology serves to classify languages based on their common features and structures, insights from linguistic typology offer translators and translation scholars an understanding of the systematic differences and similarities between languages. Crucially, such insights have proven valuable for informing translation practices and scholarship, with work such as Capelle and Loock (2016) and Teich (2003) showing clearly the importance of typological knowledge when looking at source-language translation effects. Also, in the study of translation universals, for example, linguistic typology presents a way of validating proposed universals (Molés-Cases, 2019). Contemporary contrastive linguistics, especially based on corpus data, on the other hand, provides theoretical foundations for anticipating translation problems, based on the contextualised comparison of two or more languages. Studies such as Laviosa (1998) and Ramón García (2021) argue strongly for the role of both contrastive approaches and corpus data as ways of provid-

DOI: 10.4324/9781003184454-1

ing firmer foundations for translation theory and practice. Such foundations can help researchers ensure rigor in comparison, avoid methodological circularity in research design, and offer translators and translation scholars attested examples of language use. Drawing on advances in these fields, in conjunction with the emergence of new epistemologies and approaches in translation studies itself, the field has developed apace.

Influence from parallel fields in linguistics was not to end there. Corpus linguistics, more generally, has proved revolutionary for translation studies, making particularly valuable contributions in terms of data design, collection, and analysis. The suggestion that corpora had a wide-ranging role to play in the study and practice of translation was made relatively early in the so-called corpus revolution, especially in the wake of the creation of early multilingual corpora (see, *inter alia*, McEnery & Wilson, 1994; Lauridsen, 1996; Bernardini & Zanettin, 2000). Corpus linguistics involves the analysis of carefully designed digital collections of texts, known as corpora (corpus for singular). Corpora are typically designed using a detailed sampling frame which allows for a rigorous description of the language represented in the data in terms of a range of textual and contextual factors, for example, the genre, the author, the year of production, and the source of production. Once such variables are used to define and ensure the representativeness of corpora, computational approaches are then used to analyse these representative corpora, drawing on formal features (e.g., part-of-speech) and functional use (e.g., concordance analysis).

Through this focus on form and function in richly contextualised corpora, corpus linguistics supports translation studies' concern for both the mechanics of language and contextualised language use. In turn, translation studies led corpus linguists to rethink and refine corpus-building practices, especially with regard to the key concepts of representativeness and sampling, as the reality of which texts were translated challenged ideals of corpus building (e.g., Johansson, 2007). The reciprocal relationship between corpus linguistics and translation studies has also fostered technological advances in corpus linguistics, exemplified by the development of parallel concordancing tools, specifically designed for the analysis of translated texts. While initially these tools, such as Multiconcord (King &Woolls, 1996), were only able to use small datasets and could best be viewed as specialised, nowadays they are embedded in state-of-the-art corpus searching systems, such as Sketch Engine (Kilgarriff et al., 2014), where large multilingual corpora in a wide range of languages can be searched using technology that is now mainstream, not niche. Likewise, this relationship has spurred the creation of a vast range of multilingual corpora of use for both contrastive research and for studying translation (see McEnery & Xiao, 2007, for an overview of the early years of multilingual corpus construction). Therefore, what emerges from this story is that while corpus linguistics has helped to change translation studies arguably for the better, translation studies has, in turn, changed corpus linguistics, offering new pathways for theoretical and methodological advancement.

This handbook has emerged in recognition of the myriad contributions that corpus linguistics has played in the development of contemporary translation studies. As such, the composition of the volume is equally diverse, spanning foci on the design and analysis of multilingual corpora, qualitative and quantitative approaches in corpus-based translation studies, corpus-based studies of translatorese and legal translation, and multimodal corpus-based translation studies. Ultimately, this volume attests that, by harnessing corpora, translators and translation scholars can examine language in use and consider the contextual nuances that influence translation choices at scale, and each chapter illustrates the tangible benefits and challenges of integrating empirical corpus linguistics approaches with translation theory and practice. The remainder of this chapter is written to be a foundation for the contributions that follow. With this in mind, Section 0.1.2 discusses corpus linguistics, contrastive linguistics, and translation studies with a view to clarifying the core theoretical concepts underpinning these fields, how they relate to one another, and how they relate

to the chapters that follow. Both contrastive and translation studies are considered here, as opposed to focusing solely on translation studies, owing to the importance both fields place on context and their shared theoretical concepts. This is followed by Section 0.1.3, which offers an outline of the key thematic areas addressed in this volume and a brief reflection on the future of the field.

0.1.2 Core Concepts in Corpus Linguistics, Contrastive Linguistics, and Translation Studies

While corpus linguistics, contrastive linguistics, and translation studies are distinct fields, with notably varied areas of interest, each overlaps in the context of contemporary contrastive research concerned with the comparison, in synchrony, of languages. Corpus linguistics, as mentioned earlier, is concerned with the study of language in context and, for contrastive research, comparability in corpus linguistics is determined by the sampling process used to construct corpora and the representativeness of the corpora to be compared. For both translation and contrastive studies concerned with the comparative analysis of languages, the concepts of the tertium comparationis and equivalence are key, as they afford reflexive approaches to determining and interrogating the comparability of translations and comparable language collected using a shared sampling frame. These concepts sit at the core of the chapters in this volume and, as such, are unpacked here to better contextualise them.

0.1.2.1 Representativeness and Sampling in Corpus Studies

Representativeness and sampling are cornerstone concepts in corpus linguistics, with each one mutually shaping the other. While representativeness is concerned with the language or language variety that a corpus represents, sampling pertains to the method used to select the texts that constitute that corpus (Brookes & McEnery, 2020). A sampling frame needs to be clearly defined and iteratively developed as the representativeness, and consequently the usefulness of a corpus, hinges on the quality of its sampling procedure. A complexity that emerges in the sampling process is that language is multifaceted and infinitely variable. As such, sampling strategies must be reflexive and respond to specific research aims. For example, if the goal of a research project is to study how newspaper headlines are translated, there would appear to be no need to include spoken language. However, consideration would need to be given to the types of newspapers and the types of headlines being collected, for example, tabloid and broadsheet papers, and news articles, editorials, and opinion pieces.

Likewise, it may be important to consider whether balanced or proportional sampling processes best suit the research being conducted. The former collects a balanced number of texts for the variables that form the focus of a study, while the latter determines the quantity of texts per variable based on the proportional distribution of that variable in the real population being studied. This can lead to the development of different types of corpora, such as balanced or specialised corpora (Brookes & McEnery, 2020), which, among other factors, typically differ in size, with the former generally being much larger. Ultimately, representativeness and sampling are not just methodological requirements; they are also theoretical concepts that underpin the validity and reliability of corpus linguistics research. A robust sampling strategy leads to a corpus that is a small yet accurate reflection of the language it represents, which provides a sound empirical basis for linguistic inquiry. Therefore, these concepts form a core foundation for any study using a corpus.

0.1.2.2 The Tertium Comparationis and Equivalence in Contrastive Linguistics and Translation Studies

The tertium comparationis is a fundamental concept in both contrastive linguistics and translation studies. Simply put, it is an analytical tool used to establish a common ground for comparison

between two linguistic entities (Connor & Moreno, 2005; Curry, 2023). As such, it acts as a reference point against which two or more languages are compared and, typically, the tertium comparationis is constituted by a shared linguistic feature that exists across languages. For example, to compare interrogatives in English, French, and Spanish, the tertium comparationis might be the presence of a question mark that signals a question in a text (e.g., Curry, 2021; Curry, 2024). Establishing questions as a tertium comparationis allows for further examination of how each language poses questions. However, the tertium comparationis must be established at all points of comparison to ensure a rigorous analysis. Therefore, it is not sufficient to simply identify question marks and compare them; it is also important to ensure that the texts in which the questions are posed, the contexts in which those texts were written, and so on, are also comparable. Owing to the importance attached to the tertium comparationis for ensuring comparability, contemporary comparative studies treat the tertium comparationis as both the starting and closing point of a comparison, where the assumed tertium comparationis is tested through a falsification process (Curry, 2021; McEnery & Brezina, 2022). This process of falsification is achieved through equivalence testing.

Equivalence in contrastive linguistics and translation studies serves as a theoretical device used to test the relationship between languages (Adamska-Sałaciak, 2013). Owing to terminological fuzziness in the literature, equivalence is often referred to in different ways, including as equivalent, congruence, and correspondence, for example. In comparative studies, the assumed tertium comparationis is tested in terms of equivalences, be they frequency-based, textual, contextual, syntactic, semantic, and so on. For example, to test the use of questions as tertium comparationis, equivalence testing could include determining their frequency in each language and/or their location in a text (e.g., Curry, 2021). It is possible to have more than one equivalence to test the assumption, and it is advisable to use multiple in order to test and retest the assumed tertium comparationis. Overall, the tertium comparationis and equivalence are inseparable as they mutually shape each other. They sit at the core of contrastive and translation studies and, as such, each translation study in this volume is underpinned by an assumed tertium comparationis that is analysed and tested in terms of some form of equivalence.

0.1.2.3 Corpus Approaches to Translation Studies: Key Concepts

To effectively exploit corpus approaches for translation studies, it is important to bring these theoretical concepts together. This involves consideration of multilingual corpora, the use of the tertium comparationis at each stratum of analysis, and the use of equivalence testing within a corpus linguistic approach.

Multilingual corpora are central to corpus and translation studies. These corpora are either parallel corpora or comparable corpora (Johansson, 2007), where the former pertains to collections of texts that exist as translations of the same source material in multiple languages and the latter includes comparable texts from different languages collected using the same sampling frame. For the most part, translation studies are concerned with parallel corpora, the design of which involves the alignment of texts from a source language with their translated counterparts in one or more target languages. These can be unidirectional, linking source language texts to their translations in one other language, or bi/multi-directional, involving several languages translating between each other. Such corpora have an established tertium comparationis, given that they include translations based on a directly comparable source text. As such, they are invaluable for identifying patterns in translated texts, which can provide insights into translation choices, translatorese, and the transfer of meaning across languages.

While holding obvious value for studying the language of translated texts, parallel corpora typically represent very specialised language use derived from the language produced by a small number of translators, responding to an original source text. Comparable corpora, however, offer a lens on original language use but require rigorous criteria to ensure true comparability. In the case of comparable corpora, the tertium comparationis is less obvious, and must be established through robust and iterative sampling processes and equivalence testing. As such, comparable corpora, like parallel corpora, are typically quite specialised. Both parallel and comparable corpora offer translation scholars useful data that can allow them to understand translation practices and identify effective means to translate linguistic features in different languages and contexts.

In analysing parallel and comparable corpora, a tertium comparationis must be established and tested in terms of selected equivalences. These analyses are conducted using corpus analytic approaches, and for many researchers, this involves using existing or bespoke corpus analysis software. For parallel corpora, existing software that support parallel concordancing are particularly useful, for example:

- Sketch Engine (https://www.sketchengine.eu/)
- AntPConc (http://www.laurenceanthony.net/software/antpconc/)
- ParaConc (http://www.athel.com/para.html)

Using such tools, the equivalence testing can be convergent or divergent (Chesterman, 2007), which dictates the direction of analysis. In this regard, convergent analyses focus on features assumed to be comparable across languages, while divergent analyses seek correspondences in other languages based on a feature identified in one language. Researchers must navigate and avoid methodological circularity, ensuring that the analysis does not merely confirm preconceived hypotheses but engages in a thorough falsification process. This can be done by layering and retesting the tertium comparationis from multiple directions.

By employing corpus approaches to translation studies, translators can determine empirically an effective means to reproduce a corresponding text in a target language in a way that maintains and recontextualises the meaning, style, function, and context of the source text, despite the potential for systematic differences between languages. Corpus approaches offer insight into translation practices and patterns that can inform translation decisions. Likewise, corpus approaches can offer a diagnostic resource for identifying outliers in translation practices. Moreover, corpus approaches can offer models based on authentic language use that could be used to inform and analyse translation practices. Translation is, at its core, a creative practice that requires ontological positioning and decision making on the part of the translator (Pérez-Paredes & Curry, 2024). In legal translation, for instance, function may be prioritised to ensure that the translation fulfils the same legal function as the source text, while in literary translation, translators may lean more towards preserving the aesthetic and emotional impact of the original, at the cost of other elements of the text. With the right corpus available, however, translators have a resource that can be used to inform or evaluate their practice.

0.1.3 Current and Future Directions in Corpus and Translation studies

This handbook offers comprehensive exploration of the diverse applications of corpus linguistics within the domain of translation studies, looking at the foundations of the field and new directions for development. Each chapter engages with a core element of corpus approaches to translation studies, spanning foci on data, analytical approaches, translation domains, multi-

modality, critical translation studies, and applications to teaching, and translator and interpreter training.

The volume opens with a focus on data, discussing parallel corpus design and the triangulation of multilingual corpora. Advancing on these issues, consideration is given to analytical approaches in corpus and translation studies. This includes a focus on empirical and quantitative approaches to corpus translation studies, approaches in translation and contrastive studies, translation universals and explicitation, as well as merging analytical approaches from parallel fields, including pragmatics, lexicography, cognition, and interpretation. In terms of domains of translation and linguistic features of focus, this volume addresses a range of core thematic areas in translation studies. This includes a focus on translatorese, as well the translation of texts from drama, legal, and scientific contexts. The volume also offers valuable contributions at the cutting edge of corpus and translation studies, relating to multimodality and technological advances. This includes a focus on audiovisual texts, audio description, audiovisual subtitling, and machine translation. This handbook also takes a critical perspective on key issues in translation and interpretation, addressing critical translation studies, ideology and interpreting, and gender in corpus and translation studies. Finally, the applications of corpus approaches to translation studies are addressed throughout the volume, with areas of note including the use of corpora and translation to teach language for specific purposes, the use of corpora to test translation quality, the use of corpora in translator education, and the use of corpora and translation to teach Brazilian sign language.

Corpus approaches to translation studies have heralded a shift toward nuanced, data-driven insights into the intricacies of language in context. With advancements in computational linguistics and artificial intelligence, it is likely that, in years to come, corpora will grow not only in size but also in complexity, encompassing a wider array of languages, language varieties, and specialised fields of language use. Yet, with such advances, there is a need to remain critical and cautious (e.g., Curry et al., 2024) to ensure that all developments serve the advancement of the field. The future of corpus approaches in translation studies is poised for expansion, integration, and refinement. It promises a richer, more comprehensive exploration of translation as a complex, multifaceted practice that is central to the human experience of sharing and preserving knowledge across linguistic boundaries. This volume is timely in that it allows us, as a community, to take stock of the development of the field to date, to capture the current state-of-the-art, and to identify key areas for further development.

References

Adamska-Sałaciak, A. (2013). Equivalence, synonymy, and sameness of meaning in a bilingual dictionary. *International Journal of Lexicography*, *26*(3), 329–345.

Bernardini, S., & Zanettin, F. (2000). *I corpora nella didattica della traduzione: Atti del Seminario di studi internazionale*. Bologna: CLUEB.

Brookes, G., & McEnery, A. (2020). Corpus linguistics. In S. Adolphs & D. Knight (Eds.), *The Routledge handbook of English language and digital humanities* (pp. 378–404). London: Routledge. https://doi.org/10.4324/9781003031758-20

Cappelle, B., & Loock, R. (2016). Typological differences shining through: The case of phrasal verbs in translated English. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies. New methodological and theoretical traditions* (pp. 235–259). Berlin: Mouton de Gruyter.

Chesterman, A. (2007). Similarity analysis and the translation profile. In W. Vandeweghe, S. Vandepitte, & M. Van de Velde (Eds.), *The study of language and translation* (pp. 53–66). Amsterdam: John Benjamins.

Connor, U. M., & Moreno, A. I. (2005). Tertium comparationis: A vital component in contrastive rhetoric research. In P. Bruthiaux, D. Atkinson, W. Eggington, W. Grabe, & V. Ramanathan (Eds.), *Directions in applied linguistics. Essays in honor of Robert B. Kaplan* (pp. 153–164). Bristol: Multilingual Matters.

Curry, N. (2021). *Academic writing and reader engagement: Contrasting questions in English, French and Spanish corpora*. London: Routledge. https://doi.org/10.4324/9780429322921

Curry, N. (2023). Question illocutionary force indicating devices in academic writing: A corpus-pragmatic and contrastive approach to identifying and analysing direct and indirect questions in English, French, and Spanish. *International Journal of Corpus Linguistics*, *28*(1), 91–119. https://doi.org/10.1075/ijcl.20065.cur

Curry, N. (2024). Questioning the climate crisis: A contrastive analysis of parascientific discourses. *Nordic Journal of English Studies.*

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, 4(1). https://doi.org/10.1016/j.acorp.2023.100082

Holmes, J. S. (1972). The name and nature of Translation Studies. In Holmes, J. S. (Ed.), *Translated! Papers on literary translation and translation studies* (pp. 67–80). Leiden: Brill.

Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36.

King, P., & Woolls, D. (1996). Creating and using a multilingual parallel concordancer. *Translation and Meaning*, 4, 459–466.

Lauridsen, K. (1996). Text corpora and contrastive linguistics: Which type of corpus for which type of analysis? In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994* (pp. 63–72). Lund: Lund University Press.

Laviosa, S. (1998). The corpus-based approach: A new paradigm in Translation Studies. *Meta*, *43*(4), 474–479.

McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1994). Corpora and translation: Uses and future prospects. In M. Lorgnet (Ed.), *Atti della Fiera Internazionale della Traduzione II* (pp. 246–355). Bologna: CLUEB.

McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora: What are they up to? In M. Rogers & G. Anderman (Eds.), *Incorporating corpora. The linguist and the translator* (pp. 18–31). Bristol: Multilingual Matters.

Molés-Cases, T. (2019). Why typology matters: A corpus-based study of explicitation and implication of manner-of-motion in narrative texts. *Perspectives*, *27*(6), 890–907.

Nida, E. A. (1964). *Toward a science of translating*. Leiden: Brill.

Pérez-Paredes, P., & Curry, N. (2024). Epistemologies of corpus linguistics across disciplines. *Research Methods in Applied Linguistics*, 3(3). https://doi.org/10.1016/j.rmal.2024.100141

Ramón García, N. (2021). Contrastive linguistics and translation studies interconnected: The corpus-based approach. *Linguistica Antverpiensia*, *1*, 393–406.

Teich, E. (2003). *Cross-linguistic variation in system and text*. Berlin: Mouton de Gruyter.