RiCL Research in Corpus Linguistics

# Review of Doval, Irene and María Teresa Sánchez Nieto eds. 2019. *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*. Amsterdam: John Benjamins. ISBN: 978-9-027-20234-5. https://doi.org/10.1075/scl.90

Roberto A. Valdeón
University of Oviedo / Spain and Jinan University Zhuhai / China

In *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, published in the prestigious John Benjamins' *Studies in Corpus Linguistics* series, Irene Doval and María Teresa Sánchez Nieto have gathered a selection of the contributions to the International Conference *Parallel Corpora: Creation and Applications*. The conference, held at the University of Santiago de Compostela (Spain) in 2016, focused on the exploitation of parallel corpora for diverse purposes, and more precisely for contrastive and translation studies. As the editors posit in their introduction, since the 1990s the use of corpora has changed the ways in which language and language in practice have been studied, as comparable and parallel corpora have served researchers to investigate differences and similarities between languages. Some of the first corpora (such as the *English-Norwegian Parallel Corpus* and the *English-Swedish Parallel Corpus*) had a clear academic purpose. Others have functioned as language resources widely used by researchers even though they were not the result of an academic endeavour as such, e.g. the multilingual corpora of the various European Union institutions. As can be expected, these contain institutional language and, hence, can be used for specific research purposes. But the potential of these resources, and of corpora in general, has kept growing over the past two decades. As Doval and Sánchez Nieto (3) remind us, parallel and comparable corpora are now used in machine translation and multilingual natural language processing, contrastive studies, translatology, lexicography, and also the teaching of foreign language and translation.

*Research in Corpus Linguistics* 8/1: 195–200 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.01.12

In the case of translation studies, Doval and Sánchez Nieto add, corpora have been particularly useful in applying a more empirical paradigm to descriptive studies, and, one would hope, to go beyond the many limitations of descriptivism. Many of the challenges discussed by the contributors to *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* have been recently highlighted by De Sutter and Lefer (2020: 18–19) who, in an article on a new agenda for corpus-based translation studies, have argued for

> a new multifactorial, multi-methodological and interdisciplinary research agenda for empirical translation studies […] that can potentially help us to characterize translated text, starting with linguistic features that have been said to typify other forms of constrained communication, such as non-native language varieties, editing and student writing,

moving the focus away from non-crucial parts of the corpus-based research agenda such as the study of universals (2020: 2). In addition, De Sutter and Lefer discuss corpora as process and product.

Indeed, to conclude their introduction, the editors of the book stress the two main trends in today's parallel corpora, i.e. the design and building of corpora on the one hand, and the features and applicability of corpora as products on the other. This distinction will serve to guide the readers through the well-structured contents of the collection. It is surprising, though, that no working definition of the central concept discussed in the book is provided. It is true that, although the literature abounds with definitions, it might be a mission impossible to find one that is widely used: see, for example, definitions in Olohan (2004: 24–25, 35–37) or Mikhailov and Cooper (2016: 2–8), especially in a collection of articles by various authors who are likely to use the term in slightly different ways. However, a reference to this crucial problem might serve as word of caution for the non-specialist yet interested reader.

The book is divided into three uneven sections, namely 1) background and processing, 2) creation, annotation and access, and 3) tools and application. The first section starts with a valuable introductory article on the name and nature of comparable parallel corpora. Hareide discusses a couple of definitions before providing her own "two or more parallel/translation-corpora that have the same sampling" (21), in order to underscore one of the main problems with many corpus-based translation studies, i.e. their replicability. To illustrate the usefulness of working with larger corpora defined by a number of specific parameters, Hareide uses the *Norwegian Spanish Parallel Corpus* and the *English-Spanish P-ACTRES* to test Sandra Halverson's so-called *Gravitational Pull Hypothesis*. Hareide, who testes grammatical structures, posits that the results confirm the need of "well-designed and correctly used corpora" (34).

In the next three chapters of this section, Josep Marco, Rosa Rabadán and Martin Volk offer somehow alternative uses and applications of parallel corpora. Marco defends the use of parallel corpora for two purposes: as the main source of information to analyse translators' choices and as secondary data to supplement information provided by a comparable corpus. Here we have the first difference with Hareide's combined use of 'comparable' and 'parallel' and, hence, the first example of potential confusion for the non-specialist even though, ultimately, both Hareide and Marco's efforts go in the same direction. Marco uses two examples to show the value of parallel corpora. Both draw on the *Valencia Corpus of Translated Literature* or *COVALT* to find patterns of correspondence between source and target texts when using parallel corpora, and to explain certain patterns when using comparable corpora to complement the former.

In line with Marco's chapter, Rabadán insists on the importance of combining parallel corpora with comparable and monolingual corpora. Starting with broad and narrow definitions of parallel corpora, which do not correspond exactly to those of previous chapters, Rabadán defends the need to recycle or reuse existing corpora rather than to waste time to build a new one, even if that involves upgrading existing sources to meet the demands of new research projects. Corpus efficiency can also be achieved by using comparable and monolingual corpora, or by adding new annotations to the information stored in those resources. Rabadán also includes a set of useful strategies to enhance collaborative efforts when embarking on a new project and, thus, save precious time (71). It is also worth noting that in the second section of the book, Doval *et al.* will argue in favour of creating new corpora when existing ones do not provide relevant information for specific research questions, and Sanjurjo-González and Izquierdo will claim that the creation of the *P-ACTRES Parallel Corpus* at the University of León was precisely the result of the insufficient nature of the *Cobuild* and *CREA* corpora that Rabadán and her colleagues had used until then.

For his part, Martin Volk draws on his own experience in gathering a variety of Swiss databases to emphasise the need of appropriate word alignment in parallel corpora in order to improve annotation, which in turn would be beneficial for more practical purposes, such as language learning and computational linguists wanting to evaluate the quality of automatic sentence alignment. For this purpose, standard annotations methods such as Part of Speech tagging, he claims, should be complemented with language-specific methods to deal with, for example, split verbs in German. Volk also presents prototypes that can improve word alignment and annotation and, therefore, contribute to dealing with issues such as translation error detection.

Section II comprises a total of nine chapters that delve into corpora creation, annotation and access. Some authors build upon corpora already discussed in the previous section (Molés-Cases and Oster, Sanjurjo-González and Izquierdo), while the rest introduce new ones, ranging from the intermodal corpus of *European Parliament Speeches* or EPTIC (Ferraresi and Bernardini) to the smaller *Corpus of German-Basque Literary Translations* (Sanz-Villar). In addition, some authors discuss corpora built from scratch (Doval *et al.*), while others present spin-offs from larger (e.g. Čermák on *InterCorp*, part of the *Czech National Corpus*) or different corpora (Molés-Cases and Oster on *COVALT PAR_ES*).

The section describes a number of important issues as regards creating and annotating corpora, which highlights the specific requirements of the various databases discussed. The contributors provide information on the features and challenges faced by the creators of these corpora as well as on the annotation processes. Of particular note is the article by Čermák, who underscores the amount of work required to build up a specific corpus, i.e. *InterCorp*, based at Charles University in Prague. *InterCorp* comprises texts in Czech and in other thirty-nine languages, aligned by a team of nearly 200 individuals (85). The chapters also show that the larger corpora, and subcorpora, tend to include at least one widely spoken language (typically English, French, German and Spanish), whereas lesser spoken languages (e.g. Vietnamese or Catalan in *InterCorp* (96–97), are more likely to be found in smaller corpora or subcorpora, except when the database specialises on a specific language: for example, Galician in the *CLUVI* Corpus (Gómez Guinovart), Catalan in *COVALT* (Marco) or Finnish in *PEST* (Mikhailov *et al.*). The case of *TAligner* is particularly interesting as its compilation includes German and Basque but also Spanish, as many literary texts were translated indirectly from Spanish (Sanz-Villar). Some chapters provide an innovative approach by including multimodal texts (Doval *et al.*) or what is termed as an intermodal corpus (Ferraresi and Bernardini).

As regards use, the various corpora presented in this section allow different possibilities. *InterCorp* allows users to compare up to four languages and to search one or more languages by phrase or by lemma (Čermák), *PaGeS* (Doval *et. al.*) and *PEST* (Mikhailov *et al.*) to compare dialectal variation (Doval *et. al.*), *EPTIC* to compare different communication modes (Ferraresi and Bernardini), *MULTINOT* to study contrastive differences between original texts in English and Spanish, between translations in both directions and between translated *versus* non-translated texts in both languages (Lavid López). Some have been used to produce bilingual dictionaries (e.g. *InterCorp*) or might improve computational systems in different subfields in the future (e.g. MULTINOT). Most authors stress the dynamic nature of these corpora, which allows them to set up specific goals to do research at present while also providing an opportunity to consider different objectives as the corpora evolve.

Finally, the three chapters in Section 3 cover the tools and applications of comparable and parallel corpora. Pablo Gamallo Otero discusses techniques to build highly reliable bilingual dictionaries using comparable corpora to test the validity of the choices made when creating a new dictionary by using two existing ones, and shows their value to create new dictionaries for languages with fewer resources and parallel corpora. García *et al.* also draw on Iberian languages, i.e. Spanish and Portuguese, to suggest the use of parallel corpora to extract bilingual collocation equivalents. Given the tendency by non-native speakers of a language to use unusual lexical combinations, García *et al.* stress that parallel corpora can be used to identify thousands of collocation equivalents with a very high precision (of around 86%) in an automatic and fast manner. This would allow the production of dictionaries and other teaching materials for language classroom use. It is true, however, that the success of the strategy would need to be tested with less closely related languages in order to confirm its usefulness with other language pairs. Finally, Ghoshal and Rao explore normalization processes of abbreviations and shorthand forms in French text messages. Although their experiment was successful, the objective of their work is never clearly explicated.

On the whole, the chapters in this collection make a strong case for the use of parallel, comparable, bidirectional (Lavid López; Sanjurjo-González and Izquierdo), multilingual (Čermák), intermodal quasi-parallel (Ferraresi and Bernardini) and comparable parallel (Hareide) corpora in contrastive and translation studies. They underscore their potential not only for descriptive studies but also for translator training, translation practice, machine translation, post-editing dictionary making and so on. Most of them focus on linguistic aspects (e.g. motion events in German and Spanish, adverbials in English and Catalan, Spanish gerunds and the corresponding forms in Norwegian and English, modality) that could be examined by means of corpora. But the chapters may also provide ammunition to those translation scholars who, according to Malmkjaer (1998) - quoted by Marco in this volume - feel a "disaffection bordering on hostility […] with regard to linguistics" (43). Hopefully, the arguments carefully laid out by the authors and the editors of this volume will entice some of the disaffected to the dark side as well.

## References

De Sutter, Gert and Marie-Aude Lefer. 2020. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives* 28/1: 1–23.

Malmkjaer, Kirsten. 1998. Love thy neighbour: Will parallel corpora endear linguists to translators? *Meta: Translator's Journal* 43/4: 534–541.

Mikhailov, Mikhail and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies*. London: Routledge.

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

*Reviewed by*
Roberto A. Valdeón
University of Oviedo
Department of English, French and German.
C/ Amparo Pedregal s/n
E-33011, Oviedo.
Spain
e-mail: valdeon@uniovi.es