「BOOK REVIEWS」

Understanding the state-of-the-art of parallel corpus research in the world through *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications* (Doval & Nieto, 2019)

Yasunori NISHINA

This book elucidates how parallel corpora can be compiled and utilized for contrastive studies, translation studies, and machine translation with the elucidation of the design and architecture of the parallel corpora and searching interface by the three parts of the book, including 16 chapters by the international researchers, preceded by the introduction by the two editors of the book.

The origin of parallel corpora dates back to 1971 regarding the Yugoslav Serbo-Croatian-English contrastive project by R. Filipovic. Since then, in the late 1980s, the Canadian Hansard Corpus of English and French Texts was compiled, followed by English-Norwegian Parallel Corpus (ENPC) and the English-Swedish Parallel Corpus (ESPC) in the 1990s. As Borin (2002: 1) has already stated in his book, "[i]n the last decade or so, parallel corpus linguistics has emerged as a distinct field of research within corpus linguistics, itself a fairly young discipline". Although the term 'parallel corpus linguistics' is not pervasive in the field, it has various possibilities to develop and expand the empirically-based language studies of two or more languages.

Within the 16 chapters as the main body of the contents in the book, the three chapters focus on the bilingual parallel corpora including an updated version of the ACTRES Parallel Corpus (P-ACTRES 2.0), Parallel Corpus of German and Spanish (PaGes), Corpus Linguístico da Universidade de Vigo (CLUVI), and Multidimensional Annotation of English-Spanish comparable and parallel texts for linguistic and computational applications (MULTINOT), while the four chapters deal with multilingual parallel corpora such as a part of the project Czech National Corpus (InterCorp), Valencian Corpus of Translated Literature (COVALT: the Corpus Valencià de Literatura Traduïda), the European Parliament Translation and Interpreting Corpus (EPTIC), the Parallel Electronic corpus of State Treaties (PEST), and a parallel/multilingual corpus consisting of literary texts translated from German to Basque (ALEUSKA). Most of the parallel corpora introduced in this book are text-based corpora, whilst two are multimodal (i.e. CLUVI and EPTIC). All parallel corpora consist of several text genres, text types and/or modes (e.g. written/spoken). At the same time, they are annotated at various levels.

The four chapters in the first part of the book are dedicated to how parallel and comparable corpora can usefully contribute to translation studies and contrastive linguistics, as presented in the title of the book, with the focus on some issues including the background/processing of parallel corpora and the recent developments of word-level alignment between two (or more) languages. L. Hareide's study adopted the applied use of two parallel corpora of the Norwegian Spanish Parallel Corpus (NSPC) and the first version of P-ACTRES as "comparable parallel corpora" to examine the gravitational pull hypothesis on the language pairs of Norwegian-Spanish and English-Spanish. The gravitational pull theory (Halverson, 2003, 2017) was proposed as a potential explanation for some general aspects of translated language. The underlying hypothesis is that highly salient linguistic items would be overrepresented in translational corpus data because they are more likely to be selected by translators. The NSPC and the English-Spanish P-ACTRES corpus are comparable to one another because they include similar-sized sub-corpora. With these corpora, the gravitational pull hypothesis was successfully tested, and the comparable parallel corpora method was proved useful in the Corpus-Based Translation Studies (CBTS). J. Marco's study, then, presented two case studies using the English-Catalan sub-corpus of COVALT in his chapter: the first study analyzes the translation of meal names with a parallel corpus as a main source of data, and the second analysis is on the construction *-ment* adverb + adjective as a supplementary source of data. R. Rabadán's study, on the other hand, provides her ideas about how parallel corpora can be useful and the details about the overview of resources, tools, applications, etc. This chapter firstly introduced the definition of parallel corpora, concepts of the usefulness/usability of parallel corpora, and several parallel corpora resources. Then, it reviewed the uses of parallel corpora and presented a needs analysis of parallel/multilingual corpora. Finally, it discussed whether to use the ready-made parallel corpora or build a brand-new parallel corpus, application for post-editing and assessment of translation, and useful strategies to start a new project using the parallel corpus. M. Volk's study also focuses on the annotation, alignment, and retrieval of the translation equivalents from parallel corpora, for example, giving some instances from the parallel corpus of English-German of film and TV subtitles.

The nine chapters of the book's second part present the ongoing parallel corpus project in European countries regarding the technical issues and aspects including corpus creation, annotation, and access. The current version of the InterCorp, a parallel corpus of Czech and 39 other languages compiled at Charles University in Prague, is elucidated by P. Čermák. Čermák mainly describes the size and structure of the corpus and the tools for the InterCorp, such as a corpus query tool Kontext, a text alignment editor tool InterText, and an automatic-built dictionaries of Czech-foreign languages Treq. The following chapter by I. Doval, S. Fernández Lanza, T. Jiménez, E. Liste Lamas and B. Lübke introduces the design and features of PaGes. This corpus considers the direct translation from one language to another, and does the translation direction to improve the accuracy of the contrastive analysis. The text processing, mark-up and metadata are detailed, followed by the alignment process. The functionalities of the searching tool interface for PaGes are also detailed, including the introduction of four search features, namely Fast search, User-friendly search, Multi-level search, and Display, followed by the server architecture and publishing data. In the chapter of A. Ferraresi and S. Bernardini's study, EPTIC compiled from EU parliament proceedings is detailed. The corpus includes 14 separate sub-corpora in which texts are aligned between texts and between text and video, enabling to display the actual delivery of the speech linked to the concordance lines. This is a multi-purpose corpus which currently includes three languages (English, French and Italian), from the two communication modes (spoken and written) and from the two translation modes (translated and interpreted).

X. G. Guinovart chapter then elucidates the description of the CLUVI corpus which is a sentence-level aligned parallel corpora compiled from the various genre texts of the nine languages such as Galician, Spanish, English, French, Portuguese, Catalan, Italian, Basque, and Latin. The twenty kinds of language combinations and domains are identified in the CLUVI. The corpus is human-annotated based on the TMX-based CLUVI Corpus XML specification and is available online. The VEIGA corpus is partly attached to multimedia data. The SensoGal corpus is also elucidated, which is an English-Galician parallel corpus semantically annotated using WordNet data. The following chapter by J. Lavid presents several issues that emerged from the annotation of the MULTINOT corpus, a parallel corpus of English and Spanish in both translation directions. In particular, most parts of this chapter detail the annotation of semantic, pragmatic, and discourse phenomena.

The chapter by M. Mikhailov, M. Santalahti and J. Souma describes PEST. PEST is a parallel corpus of treaties concluded between Russia-Finland, Finland-Sweden, and Sweden-Russia with a sub-corpus of international conventions in all the three languages as reference data. Among all, the structure of the sub-corpora of PEST is detailed, including the number of treaties over time, the topic of treaties, and so on. T. Molés-Cases and U. Oster's chapter introduces the architecture, compilation, and indexation of the part of COVALT corpus, a parallel corpus compiled from novels in English, French, and German as source languages and Spanish as a target language. Since the authors made it analyzable online with CQPweb, the sample query with CQPweb is also visually illustrated with some language syntax and CQP syntax options. H. Sanjurjo-González and M. Izquierdo describe P-ACTRES 2.0, an extended version of P-ACTRES 1.0 adding a new sub-corpus of original Spanish texts and their English translations. The workflow of building this corpus is well described such as compiling, formatting, aligning, tagging (with Treetagger), and indexing texts. For instance, formatting texts are separately explained such as cleaning texts, transforming them into XML format, validating XML files, and carrying out a sentence division. The usability of this corpus is also well presented through the instance of Spanish translational options of English construction *with* + NP + *-ing*. Finally, in Z. Sanz-Villar's chapter, the overview of the Basque corpora and the Aleuska corpus is described. The Aleuska corpus is a trilingual corpus of German literary texts with their translations of Basque and Spanish. The corpus design and compilation process using TAligner 3.0 are presented, followed by lemmatization and annotation at the POS level. Finally, the process of the extraction of Basque multi-word expressions (MWE) consisting of onomatopoeia and a verb is presented.

The three chapters in the third part of this book deal with the tools and applications of parallel corpora concisely from the three case studies. The study by P. Gamallo successfully presents that comparable corpora can be an alternative option to generate bilingual lexicons. The study introduces the two approaches of transitivity through intermediary dictionaries and refers to the similarity of bilingual cognates. It shows that accuracy is not much different when using a comparable corpus than when using a parallel corpus. The second chapter of the final part is the study by M. Garcia, M. García-Salido, and M. Alonso-Ramos about extracting bilingual collocation equivalents from parallel corpora. They utilize dependency parsing to extract the candidates of monolingual collocation and use the bilingual model of distributional semantics to identify the equivalents of the base and the collocate of the monolingual collocations. In the study procedure, the authors focus on lemmas instead of tokens regarding both the monolingual collocations and the distributional model. The cosine distance of the vectors of the two words is calculated to determine their semantic similarity. Using the parallel corpus of normalized and non-normalized French text messages, the last chapter by P. Goshal and X. Rao observes the efficacy of the two tools/approaches of *multivec* (multilingual word embeddings) and *moses* (character-based machine translation) for text normalization. The study concludes that *moses* and *multivec* differ in their preference for normalizing different categories. As a result, the authors suggest that the two approaches should be combined for a more robust precision and result in the normalization of shorthand forms in text messages.

As a whole, this book comprehensively introduces the development of parallel corpora and search tools in European countries at the present time, which makes it a valuable book for researchers, specialists, practitioners, and graduate students in the field. However, the volume of each part was noticeably different, and we would have liked to see improvements in this respect. In particular, since the overall focus was on the technical or technological aspects of parallel corpora, a few more specific case studies of CBTS and contrastive linguistics could have been introduced. For instance, a more concrete case study of computer lexicography would have conveyed more about the operational usefulness of parallel corpora and the potential for research development. To add, it would have been helpful to introduce parallel corpus studies of languages other than European languages, such as Asian languages.

Through this book, it can be also found that there are several limitations in the parallel corpus studies at the moment due to the limitation of bilingual/multilingual corpora available and the specificities of those corpora, as compared to the monolingual corpora. For instance, the size of EPTIC is still small, and only three languages are available out of 23. This situation is applied to other parallel corpora as well. The spoken parallel corpora are particularly scarce in the corpora of the certain pairs of languages (e.g. Japanese-English pair, see Nishina, 2023). The balance of translation direction is also important in parallel corpus studies. With some exceptions including MULTINOT, however, many parallel corpora are unidirectional.

In addition, the multimedia parallel corpora are scarce compared to the text-only corpora due to the scarcity of the translated multimedia materials. As X. G. Guinovart pointed out, its further development is strongly expected since the multimedia parallel corpora can be helpful for educational purposes, facilitating learners' autonomous language learning and providing meaningful information about ready-made subtitles to translators/practitioners.

Another issue in compiling parallel corpora is the copyright issue, whether most of the documents planned to be included in the corpora are publicly available. In the case of Japan, several Japanese-English parallel corpora are available, although a paid license agreement must be signed for the use of some of them (Nishina, 2023). For this reason, the development of parallel corpora incurs significant research costs.

Furthermore, the limited search tools and their functions available may be a problem. This is because the progress of tool development and its performance varies from language to language. There is a need for a unified platform on which researchers worldwide can share and collaborate to develop tools for bilingual/multilingual corpora. To add, as a parallel corpus linguist, I tackle the compilation of Japanese-English and English-Japanese parallel corpora and their searching interface, and have faced a similar situation as presented in this book; the sole researcher or a few members of a team feel that there is a limitation to developing and expanding the parallel corpus resources, interface, and researches. Co-operation with the international researchers/teams that have and provide the diverse expertise, skills, and resources is necessary for the near-future parallel corpora linguistics as easily expected. For instance, my team developed the lexical profiling system of the several Japanese-English parallel corpora available, named Parallel Link (https://www.parallellink.org/), and this tool enables any corpus users to extract the linguistic information they need quickly in terms of patterns, collocations, and instances with their translation equivalents. We hope to expand this tool into multilingual ones including, for example, Spanish. However, modern languages such as Spanish, French, and German are outside our field, and European researchers' assistance is required to succeed in this future project.

Finally, a general reference parallel corpus does not yet exist. Suppose there is a general reference parallel corpus. In that case, it can be utilized for the extraction of meaningful translation units, a compilation of bilingual dictionaries, empirical-based bilingual studies including translation studies/teachings, material developments, and language processing. As represented by British National Corpus (BNC) and Corpus of Contemporary American English (COCA), I sincerely hope that a general reference parallel corpus will be created in the near future.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 20K00692 and 23K00599. I would like to express my gratitude to the anonymous reviewer for his/her helpful comments and suggestions.

References

Borin, L. (2002). …and never the twain shall meet? In L. Borin (Ed.), *Parallel Corpora, Parallel Worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999* (pp. 1–43). Rodopi.

Halverson, S. (2003). The cognitive basis of translation universals. *Target, 15*(2), 197–241. https://doi.org/10.1075/target.15.2.02hal

Halverson, S. (2017). Gravitational pull in translation: Testing a revised model. In G. De Sutter, M. A. Lefer, and I. Delaere (Eds.), *Empirical Translation Studies: New Methods and Theoretical Traditions* (pp. 9–45). Mouton de Gruyter. https://doi.org/10.1515/9783110459586-002

Nishina, Y. (2023). *Aspects of Parallel Corpus Linguistics: From Monolingual Corpus Studies to Bilingual Corpus Studies (Society of Global Communication Studies of Kobe Gakuin University Research Series Vol.2)*. Kaitakusha.

（仁科　恭徳　神戸学院大学）