RiCL
Research in Corpus Linguistics

Review of Julia Lavid-López, Carmen Maíz-Arévalo and Juan Rafael Zamorano-Mansilla. 2021. *Corpora in Translation and Contrastive Research in the Digital Era*. Amsterdam: John Benjamins. ISBN: 978-9-027-20918-4. DOI: https://doi.org/10.1075/btl.158

Mikhail Mikhailov
Tampere University / Finland

1. INTRODUCTION

The book is a collection of articles written by authors from Spain, Great Britain, Germany, Switzerland and other countries. The papers are based on the presentations held at the *International Symposium PaCor* 2018 (*Parallel Corpora: Creation and Applications*, Madrid, November 2018) hosted by the research group FUNCAP[1] in collaboration with the Institute of Modern Languages and Translation and members of the Department of English Studies at the Complutense University of Madrid (UCM).

The papers present research in the field of corpus-based translation and contrastive studies. The authors work with different pairs of languages: English and Spanish, English and German, English and Chinese, English and Portuguese, English and Turkish, and even Old English and Modern English. All the papers deal with parallel or comparable corpora.

The usefulness of multilingual corpora in contrastive and translation studies has been promoted by many researchers starting in the 1990s (see, e.g., Baker 1995; Johansson 2007; McEnery and Xiao 2008). However, multilingual corpora have traditionally got less attention than monolingual corpora (see, e.g., Kenning 2010;

---

[1] https://www.ucm.es/funcap/el-grupo

Mikhailov and Cooper 2016: 1–2). The first conference devoted to parallel corpora was organized in Uppsala, Sweden, in 1999 (Borin 2002). Then, after a fifteen-year-long break, the topic was resumed at the *PaCor* 2016 symposium at the University of Santiago de Compostela (Doval and Sánchez-Nieto 2019). More and more publications have appeared on this subject (see, e.g., Bernardini 2011; Frankenberg-Garcia 2009; Tiedemann 2012, among others), yet one cannot claim that no stone has been left unturned.

The book is divided into two parts. In the first part ("Corpus resources and tools") the issues of collecting and querying corpora are discussed. The second part ("Corpus-based studies and explorations") consists of case studies based on findings from parallel and comparable corpora.

2. SUMMARY

The introductory chapter by Julia Lavid-López is not only a guide of the volume (as often happens), but also explains the idea of the book, which is to show the scope of available data and to introduce the tools that can be used for its querying.

The chapter begins with a solid overview of the corpus resources with an emphasis on parallel corpora. The author does a brief historical tour which is very helpful for the readers with little background in the field. The most prominent and important projects are mentioned: *English Norwegian Parallel Corpus* (ENPC),[2] ACTRES,[3] *The European Parliament Proceedings Parallel Corpus* (Europarl),[4] *Multilingual Text Tools and Corpora* (Multext),[5] *The Open Parallel Corpus* (OPUS),[6] and CLARIN ERIC.[7] The chapter also introduces the main challenges of compiling parallel corpora: limited availability of parallel texts from certain domains, genres, time spans, and for certain pairs of languages, as well as imbalance in the direction of the translations.

The next section is devoted to corpus-related tools: Translation Memory (TM) systems and corpus management tools. Personally, I would not have assigned TM systems

---

[2] https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/
[3] https://actres.unileon.es/wp/
[4] https://www.statmt.org/europarl/
[5] https://cordis.europa.eu/project/id/LRE62050
[6] https://opus.nlpl.eu/
[7] https://www.clarin.eu/

to corpus software. They are used for entirely different purposes: facilitating and automating translation process (Computer Aided Translation tools). Corpus-related functions, like parallel concordancing, are add-ons, and *Trados*[8] or *WordFast*[9] concordancing is much less flexible than in real corpus management systems like *Sketch Engine* (Kilgarriff *et al*. 2014) or the *IMS Open Corpus Workbench* (CWB).[10] However, there is some logic in introducing TMs together with corpus tools, because TM technology is little by little catching up with corpus technologies, and, as we will see, one of the chapters (see Ranasinghe *et al*. below) deals with making TM more intelligent. The overview of corpus management tools demonstrates that they are still very much oriented on monolingual corpora. Most of the tools mentioned in the chapter were initially developed for monolingual corpora and have additional functionality for querying parallel corpora as well. The current developments include the constantly growing role of web-based software and extensive use of Corpus Query Language (CQL) querying.[11] The most popular tool is *Sketch Engine*, and it is not only a research tool but is quite suitable for practical tasks, like copyediting or translating.

The chapter shows that the development of parallel corpora and corpus tools should serve both contrastive studies and translation studies and, at the same time, can be available for translation practitioners.

2.1. Part I: Corpus resources and tools

The first chapter in Part I ("A fresh look at language technologies and resources for translators and interpreters") by Gloria Corpas Pastor and Fernando Sánchez Rodas provides a brief outline of IT-resources for translators and interpreters. The authors point out that cardinal changes have took place in translation process. These are expansion of Computer Aided Translation (CAT) tools and Neural Machine Translation (NMT), cloud technologies, and crowdsourcing. Post-editing machine-translated texts becomes a routine, not an occasional task. The 'traditional translation' is being rapidly displaced. Although the field of interpreting is more conservative towards technologies and, at the moment, is still falling behind translation, it is also experiencing significant changes.

---

[8] https://www.trados.com/products/trados-studio/
[9] https://www.wordfast.com/
[10] https://cwb.sourceforge.io/
[11] https://www.sketchengine.eu/documentation/corpus-querying/

Different kinds of remote interpreting have become part of everyday life, and on-site events with interpreter online (telephone-mediated interpreting, video-mediated interpreting) are being replaced by cloud events with all participants communicating online via teleconferencing. As a result, the cloud-based interpreting is experiencing a fast growth.

The authors point out that, in spite of many advantages they give to translators, CAT tools still have many weak points and do not provide optimal solutions in many cases. Text corpora can in many cases complement CAT and Machine Translation (MT) and assist translators in many tasks. The main advantage of the corpora is the availability of huge amounts of data. Using corpora in interpreting is less obvious, yet there is some development here as well. Corpora are used by interpreters mainly in the preparation phase, and there also exist interpretation corpora that are used in interpreting studies and for interpreter training.

Another type of tool mentioned in the chapter is computer-assisted interpreting (CAI). However, these are multi-purpose tools, among them digital pens, note-making tools, and terminology management tools. These instruments are designed for a large group of users including interpreters.

Currently, translators have significantly more tools at their disposal than interpreters. Among translators, the most active users of CAT tools and MT are working in the field of specialized translation. Literary translators usually reject these tools but refer favorably on corpora that help them in looking up better equivalents, translation solutions, or check usage of a certain word or phrase.

The authors claim that the four stages of machine translation acceptance defined by Sgourou (2019; (1) nescience, (2) contempt, (3) reluctant adoption and shame, and (4) acceptance) are applicable to acceptance of all kinds of technological innovations in translating and interpretation. Translators are now in stage (4), while interpreters are somewhere in between stages (3) and (4).

Actually, the authors of the chapter present two different kinds of technologies: those supporting technical processes of translation and interpretation (scheduling, data sharing, teamwork, transmission) and those supporting language services (checking lexical units and terminology, grammar check, looking up translation equivalents, etc.). It would have been better to deal with them separately and to point out the differences

between the utilities designed especially for translators (e.g., CAT tools), for a wide range of language service providers (e.g., corpora), and for all users (Optical Character Recognition, MT).

Chapter 2 by Yi Gu and Ana Frankenberg-Garcia ("ZHEN: A directional parallel corpus of Chinese source texts and English translations") is devoted to parallel corpora with the language pair of English and Chinese. The authors point out that most of existing parallel corpora of this language pair are collections of translations from Chinese into English, a large number of them being translated by native speakers of Chinese with post-editing by English native speakers. Although the existing corpora contain a certain amount of Chinese-English translations, it is difficult to detect those because the source text is not specified for official documents, as is the case with United Nations (UN) texts. The English-Chinese parallel texts are usually more difficult to obtain, and many of them are old texts from the nineteenth century.

In the chapter, a new English-Chinese corpus is introduced. The authors outline the criteria for selecting texts for the corpus, existing difficulties in searching and collecting parallel texts, and the technique for looking up source texts and translations. The resulting corpus represents various text genres, such as government documents, white papers, UN documents, fiction, political speeches, movie subtitles, academic abstracts, etc. The source texts are written in Mainland Mandarine Chinese and are published after 1990 (with few exceptions). The corpus was compiled with *Sketch Engine* and can be shared with other researchers. The authors show the advantages of the resource compared to other English-Chinese datasets and outline its possible uses.

In Chapter 3, "Word alignment in a parallel corpus of Old English (OE) prose. From asymmetry to inter-syntactic annotation," Javier Martín-Arista presents a parallel corpus with Old English texts and their translations into the Present-day English (PDE), with multiple examples that demonstrate morphological and syntactic differences between OE and PDE. This type of corpus is not very common, and it has certain technical issues that need to be solved. The *Open Access Annotated Parallel Corpus Old English* (ParCorOE)[12] targets 300,000 running words, and consists of OE texts of various genres and their translations into PDE. The collection is fairly large for this type of corpora. The texts are aligned at sentence and word levels, lemmatized, and include morpho-syntactic

---

[12] https://www.nerthusproject.com/search-parcoroe

annotation. The syntactic structures can be visualized as graphs. The corpus is freely accessible online.

Chapter 4 by Tharindu Ranasinghe, Ruslan Mitkov, Constantin Orăsan and Rocío Caro Quintana is entitled "Semantic textual similarity based on deep learning: Can it improve matching and retrieval for Translation Memory tools?" Current TM tools are based on string matching techniques (Levenshtein's distance, Dice-Sørensen index; see Levenshtein 1966 and Sørensen 1948, respectively). These methods work well on pairs of sentences which are lexically and syntactically close. Using semantic similarity helps to find sentences with other lexemes and/or other grammatical constructions used to express the same meaning. Semantic similarity measures are distance measures of semantic vectors of sentences which are the result of pairwise comparing sentences from large datasets. In this chapter, the authors try to find out whether using semantic textual similarity has perspectives. They test various semantic sentence encoders (InferSent,[13] Universal Sentence Encoder,[14] and SBERT[15]) and compare the results with Okapi[16] which uses Dice-Sørensen index. The testing is done on the English-Spanish Directorate-General for Translation of the European Commission (DGT) TM. The results show that the semantic encoders are fast enough to be used in industry and that they are more efficient with the sentences with low string similarity. They also produce less bad matches resulting of partial coincidence of the sentences.

In Chapter 5, "TAligner 3.0: A tool to create parallel and multilingual corpora," Zuriñe Sánz-Villar and Olaia Andaluz-Pinedo introduce a tool for working with parallel corpora. Unlike many existing corpus tools, *TAligner 3.0*[17] is an open source and cross-platform tool that can align multiple translations of the same text and even retranslations. The software has also special features for aligning dramatic texts. The search routines provided are frequency lists and parallel concordancing.

The tool belongs to the third generation software, that is, it works on workstations, and not on servers. The authors are aware that this creates certain problems with

---

[13] https://github.com/facebookresearch/InferSent
[14] https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder
[15] https://www.sbert.net/
[16] https://okapiframework.org/wiki/index.php/Main_Page
[17] https://addi.ehu.es/handle/10810/42445?locale-attribute=en

installations on local computers and sharing corpora (especially in the case of large corpora).

Chapter 6, "Developing a corpus-informed tool for Spanish professionals writing specialised texts in English," by María Pérez Blanco and Marlén Izquierdo demonstrates direct practical applications of multilingual corpora. *Promociona-TÉ*, a tool for generating product descriptions for the tea industry, was a result of cooperation between the ACTRES research group and a tea manufacturer Pharmadus Botanicals, S.L.[18] The tool is based on the data from a comparable English-Spanish corpus. The instruments of this kind are very important for small enterprises which cannot afford commissioning translators, and for which machine translation of specialized texts does not yield sufficient quality because of the limited availability of parallel texts.

2.2. Part II: Corpus-based studies and explorations

In Chapter 7 ("English and Spanish discourse markers in translation: Corpus analysis and annotation"), Julia Lavid-López presents an analysis of the English discourse markers (DM) *in fact*, *actually*, and *really* and their Spanish equivalents. The author uses large parallel corpora from the OPUS corpus collection available at *Sketch Engine*. The Spanish translation correspondences of the three DMs are first collected from English-Spanish parallel concordances. The corpus provided the most typical Spanish translation correspondence for the DMs in question. The analysis of data also makes it possible to define the meanings of both English and Spanish DMs. The chapter contains many usage examples and interesting findings on usage, meanings, frequencies, and interrelation of the markers.

The OPUS datasets provide large amounts of data which helps to find the most typical pairs of equivalents and perform quantitative analysis. However, the information on direction of translation is not available and it is therefore not possible to define subcorpora with original English texts and their Spanish translations, and with original Spanish texts and their English translations. Another problem is that source texts can be written by non-native speakers, and also some translations can be performed by non-native speakers of the target language. Finally, a translator may misunderstand the text

---

[18] https://www.pharmadus.com/

and use a wrong DM as equivalent. These issues can influence the use of DMs in both languages (see also Mikhailov 2021). However, large, clean and reliable parallel corpora are still to be acquired, and results obtained from noisy data are also valuable.

In Chapter 8, "The discourse markers *well* and *so* and their equivalents in the Portugese and Turkish subparts of the TED-MDB corpus," Amália Mendes and Deniz Zeyrek continue the discussion on DMs in original texts and their translations. The researchers study cross-lingual correspondences of the English DMs *well* and *so*. The data used comes from TED Talks transcripts. The TED Talks presentations are transcribed by volunteers and translated by other volunteers into other languages. The data is freely available, thus providing a valuable multilingual dataset. The TED MDB corpus (see Zeyrek *et al.* 2020) is compiled of such parallel texts with discourse relations annotated. The information on discourse relations is still not available from large corpora, and most research is therefore still carried out on small data. The research demonstrates that discourse markers of the source text are often omitted in translations, but the tendencies are different for different language pairs. The English marker *well* is sometimes kept by Portuguese translators, while Turkish translators leave it out. At the same time, the marker *so* is usually left out in Portuguese translations and often preserved in Turkish talks.

Although the methodology is interesting and promising, more data would be needed. The size of the corpus is less than 20,000 running words with about 7,000 tokens per language. The case studies are carried out on 12 examples of *well* and 30 examples of *so*.

In Chapter 9, "Variation of evidential values in discourse domains: A contrastive corpus-based study (English and Spanish)," Juana I. Marín-Arrese studies evidentiality, that is, marking the source of information, in oral and written communication in English and Spanish. The research is carried out on comparable corpora. Two types of evidentiality markers, indirect-inferential evidentiality (IIE) and indirect-reportative evidentiality (IRE), are compared. The study demonstrates that IIE markers are more extensively used both in English and in Spanish, and that the use of evidentiality markers is different in oral and written language. The data from both languages demonstrate the same tendencies with some minor differences.

Chapter 10 ("Translation for dubbing of Westerns in Spain: An exploratory corpus-based analysis") by John D. Sanderson, presents an analysis of the lexis in the American westerns dubbed into Spanish. The study is based on a parallel corpus which, at the

moment of publication, included transcripts of 20 American westerns from 1939 to 2012 aligned with the transcripts of the films dubbed into Spanish. The author studies the impact of the censorship of Franco's dictatorship on the choice of equivalents. For example, the culture-bound word *marshal* tends to be domesticated in earlier films (*comisario*, *alguacil*) and foreignized in later films (*sheriff*); rude expressions like *son of a bitch* are avoided: although there exists an exact match in Spanish (*hijo de puta* 'son of a whore'), an artificial *hijo de perra* ('son of a she-dog') is used (although the word *puta* 'whore' is nevertheless used as a separate lexeme). The study points out that a special sociolect for translating American westerns has been developed and some equivalents are still being used even now, many years after Franco's decease. In the chapter the practical use of parallel corpora of film transcripts for translating is also mentioned.

In Chapter 11, "Generic analysis of mobile application reviews in English and Spanish: A contrastive corpus-based study," Natalia Mora López explores the composition of texts in the genre of online review. She compares English and Spanish reviews from *Google Play Store*. The data is a small corpus of 200 texts (100 English, 100 Spanish) drawn from a larger text collection. The study is based on the Appraisal Theory, which aims at detecting positive and negative attitudes expressed in texts. A number of patterns are found and their features studied. In many cases the attitudes can be detected on the lexical level, although some texts, especially spam, can be misallocated. No significant differences between English and Spanish reviews were found.

In Chapter 12 ("Exploring variation in translation with probabilistic language models"), Alina Karakanta, Heike Przybyl, and Elke Teich compare the language of translations and interpretations in relation to the language of original written texts and speech. The data used are obtained from the Europarl-UdS corpus, with written texts and translations originated in the European Parliament, and several interpreting corpora. The target languages are English-German and English-Spanish translations and interpretations. The metadata of the corpora make it possible to select the data produced by native speakers both for source texts/speeches and for translations/interpretations.

The method used is Kullback-Leibler Divergence (KLD), which allows to measure probability disruptions in the data being compared. The findings are visualized as word clouds. The word probabilities are compared pairwise for translations vs. originals, interpretations vs. originals, and translations vs. interpretations. The method allows to

detect the words typical of a certain type of data, such as for German original texts as opposed to texts translated from English into German, etc. The results demonstrate that the language of translations and interpretations differs from that of the texts/speeches originally produced in the same language. Possible reasons include differences in the process of creating original text/speech and translation/interpretation, as well as the 'shining through' of the source language in translations/interpretations. However, some similar effects are detected in German and Spanish data, which demonstrates that not all features can be interpreted in terms of 'shining-through'.

In Chapter 13 ("Binomial adverbs in Germanic and Romance languages: A corpus-based study"), Johannes Graën and Martin Volk present a method of extracting binomial adverbs (*more or less*, *here and now*, etc.) from large multilingual corpora. The study is performed on the *large-Scale PARallel Corpora to study LINGuistic variation* (SPARCLING)[19] and includes six languages: English, French, German, Italian, Spanish, Swedish. Detecting multiword expressions, and binomial expressions among them, is very important both for linguistic research and for automated language processing (parsing, MT). Direct queries like Adv + Conj + Adv do not have enough recall because of the parsing errors; therefore, lists of adverbs based on morphological annotation were extracted from the corpus and the searches performed on these lists after their cleaning up. The candidates were filtered out using MI-scores and the boundaries of the multiword expressions were checked with the help of entropy values. The interlingual correspondences in parallel corpora worked as additional criteria for detecting binomial adverbs.

3. DISCUSSION

The book addresses researchers working in the fields of translation studies, contrastive studies, and corpus linguistics. Some of the papers in the first part deal with the issues of language technologies, and many papers from the second part are connected with discourse analysis. The book shows well enough the state of the art in the field: the studies presented use different methods and approaches and are performed on data of very different nature. The volume also reveals the main tendencies in modern corpus research:

---

[19] https://www.cl.uzh.ch/en/texttechnologies/research/corpus-linguistics/sparcling.html

the size of datasets is growing, new languages and language pairs are being studied with the help of corpora, new kinds of data (e.g., interpretation corpora) are being collected, descriptive statistics is being replaced by sophisticated quantitative methods, etc. It also becomes clear that the data available is not sufficient for all kinds of research and that automated annotation has many weak points.

All chapters present original research and fit well into the composition of the book. Although the studies are devoted to different language pairs, they are focused on methodological issues rather than on findings in particular languages and therefore all of them are of interest for researchers working with other languages.

As it often happens with conference volumes, the chapters are written by different authors and present different topics, which makes the book rather heterogeneous. Some papers are very easy to read and are more practically oriented, others make use of complicated methods and need more effort. Still, the volume does not require special background. All in all, the book is a suitable reading for someone interested in multilingual corpora and their use in contrastive and translation studies. It will hopefully inspire more research in the field.

REFERENCES

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

Bernardini, Silvia. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS – A Journal of Professional Communication* 26: 2–13.

Borin, Lars ed. 2002. *Parallel Corpora, Parallel Worlds. Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*. Amsterdam: Brill.

Doval, Irene and María Teresa Sánchez Nieto. 2019. Parallel corpora in focus: An account of current achievements and challenges. In Irene Doval and María Teresa Sánchez Nieto eds. *Parallel Corpora: Creation and Applications*. Amsterdam: John Benjamins, 1–15.

Frankenberg-Garcia, Ana. 2009. Compiling and using a parallel corpus for research in translation. *International Journal of Translation* XXI/1: 57–71.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: John Benjamins.

Kenning, Marie Madeleine. 2010. What are parallel and comparable corpora and how can we use them? In Michael McCarthy and Anne O'Keeffe eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 487–500.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1/1: 7–36.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8: 707–710.

McEnery, Tony and Richard Xiao. 2008. Parallel and comparable corpora: What is happening? In Gunilla Anderman and Margaret Rogers eds. *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters, 18–31.

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

Mikhailov, Mikhail. 2021. Mind the source data! Translation equivalents and translation stimuli from parallel corpora. In Vincent X. Wang, Lily Lim and Defeng Li eds. *New Perspectives on Corpus Translation Studies*. Singapore: Springer*,* 259–279.

Sgourou, Maria. 2019. The four stages of machine translation acceptance in a freelancer's life. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019), Varna, Bulgaria*. Shoumen, Bulgaria: Incoma Ltd., 134–135. https://aclanthology.org/W19-8700/

Sørensen, Thorvald. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab* 5/4: 1–34.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC' 2012). Istanbul, Turkey: European Language Resources Association, 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Zeirek, Deniz, Amália Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources & Evaluation* 54: 587–613.

Reviewed by
Mikhail Mikhailov
Languages Unit, ITC
Tampere University
FI-33014 Tampere University, Finland
e-mail: mikhail.mikhailov@tuni.fi