Mikhail Mikhailov & Robert Cooper. 2016. *Corpus linguistics for translation and contrastive studies: A guide for research.* Routledge Corpus Linguistics Guides. London & New York: Routledge. Pp. 234.

Reviewed by Olli O. Silvennoinen

Corpus linguistics is by now an established method even in fields that require comparable and/or parallel data on multiple languages, such as translation studies, contrastive linguistics and language typology. Despite this, introductions to corpus linguistics are heavily biased towards monolingual corpus linguistics in general and English corpus linguistics in particular. The present volume is a successful attempt to close this gap.

The book is published in a series of guidebooks on corpus linguistics. This shows in the contents and design of the book, which are more practical than in most introductions to corpus linguistics (e.g., Biber et al. 1998; McEnery & Hardie 2011), though not as hands-on as Gries (2009), for instance. The book mostly teaches by example, with many of the chapters consisting predominantly of case studies. The case studies are largely based on the authors' own research, which is reflected in the fact that the vast majority of them contrast Finnish with either English or Russian (Mikhailov is a professor of Russian-Finnish translation at the University of Tampere, and Cooper has worked at the English department of the same university). However, they are generally written in a way that should make them easy to understand even for readers who do not know any Finnish or Russian. In addition to the seven chapters, the book contains a foreword, final remarks, a glossary, two appendices and an index. References are given at the end of each chapter rather than at the end of the book.

The authors cite Teubert's (1996) classification of multilingual data in corpus linguistics into parallel, comparable and translation corpora (p. 5). Parallel corpora are corpora that consist of source texts with their translations that have been aligned at the level of words, sentences, paragraphs or whole texts. Comparable corpora are datasets in different languages that are extralinguistically similar (e.g., novels in English and French). Translation corpora include translated texts; their purpose is usually to study the properties of translations (*translationese*) in comparison with non-translated texts. Mikhailov and Cooper explicitly state that their focus will lie on parallel corpora. Given the technical and practical challenges involved in compiling and using parallel corpora, this

is an understandable choice that indeed makes the book stand out from other introductions to corpus linguistics. As a result, however, the book does not quite cover all the ground that its title might lead one to assume. For instance, the "combined corpus approach" (Mortier & Degand 2009), which makes use of both parallel and comparable data, might have been useful for those readers embarking on research projects in contrastive linguistics. Similarly, translation studies also uses corpus linguistics to uncover features of translationese or translation universals using translation corpora (see papers in, e.g., Mauranen & Kujamäki 2004). While these issues are touched upon, the use of comparable and translation corpora in translation studies and contrastive linguistics is largely outside the scope of the book.

However, what the book loses in coverage, it gains in coherence. The steps of each analysis are presented clearly, and as a result the reader gets a well-rounded picture of what parallel corpora are about in translation studies and contrastive linguistics and what is special in them when compared to traditional monolingual corpora.

Chapter 1 ("Parallel text corpora: a general overview") covers basic issues in corpus linguistics and the use of parallel corpora. It considers such issues as the different types of corpus and the problems in using a corpus. The authors also devote a section to the use of corpora in translation and conclude that building parallel corpora is beyond what most translators would want to invest their time in. The chapter presupposes some corpus-linguistic terminology, which may not be familiar to absolute beginners in the field, such as *type*/*token ratio*, *collocate* and *concordance*.

Chapter 2 ("Designing and compiling a parallel corpus") walks the reader through the stages of compiling a parallel corpus, from planning the corpus design and inputting the texts, through aligning and annotating the texts, to storing the corpus. It ends with a discussion of copyright issues relevant to corpus linguistics. This is a highly useful chapter that contains a lot of essential information and food for thought even for those who use a ready-made corpus.

Chapters 3, 4 and 5 are the heart of the book. They offer lucid and interesting examples of the basic ways of doing research in corpus linguistics. The chapters are ordered from the most elementary to the most advanced. Chapter 3 ("Using parallel corpora: basic search procedures") covers the very basics of corpus-linguistic research, with sub-sections devoted to conducting corpus searches, concordances, frequency lists and collocations. As well as showing how to perform these methods with

parallel corpora and what they are used for, the chapter introduces basic concepts in corpus linguistics, such as *precision* and *recall*, *n-grams* and *KWIC*. On reading the chapter, even a novice should have a good idea of what the basic procedures are and why they matter. The only problem that I would like to raise is that precision is defined in a misleading way: the definition given by the authors suggests that precision is the proportion of false positives, while actually it is the proportion of true positives.

Whereas Chapter 3 covers ground that is common to all textbooks of corpus linguistics, Chapter 4 ("Processing search results") moves to issues that are specific to parallel corpora, taking the basic search procedures one step further. It consists of four sub-sections, each of which is structured around a case study. The first sub-section concerns comparing translation equivalents in parallel concordances, probably the first thing most people would use parallel corpora for. The case study concerns the Russian adverb *pravda* 'actually, really' and its translation equivalents in Finnish. The reader is shown the process of querying the data, removing noise and categorising the tokens. The section also considers the possible effect of translators' preferences for given equivalents. The only problem with the section is that the Finnish translation equivalents are not translated into English in the running text, which may make the discussion somewhat hard-going for readers who are not proficient in Finnish and/or Russian (a rare problem in the book, which generally manages to convey the meanings of Finnish and Russian data quite well). The second sub-section shows how a similar study may be done using frequency lists as the starting-point. The case study for this section concerns the English verbs *say* and *tell* and their Finnish translations *sanoa* and *kertoa*. The authors show that genre-based translation preferences can be discovered using frequency lists rather than concordances.

The third sub-section of Chapter 4 moves on to a more fine-grained analysis by considering collocations. This time, the case study is on the English adjective *clear* and its Finnish equivalents *kirkas*, *selkeä* and *selvä*. The head nouns of the Finnish adjectives are categorised according to semantic domain, which reveals patterns of usage that are not often captured even in monolingual dictionaries. The fourth and last sub-section concerns the seemingly incongruous topic of parallel corpora in monolingual studies. Using English *before* as illustration, the section shows how the French translation of the word may be used for teasing apart locative uses from temporal ones since French makes a lexical distinction between the two (*devant* for locatives, *avant* for temporals). The use of

locative *before* instead of *in front of* appears to be highly context-sensitive, as body parts (*before my eyes*) and archaic genres such as legislative texts (*before the jury*) strongly favour its use. Appropriately enough for a textbook, this highlights the fact that the selection of data is of paramount importance in corpus linguistics. Legislative texts are often used in parallel corpus studies because of their easy availability even though they might not represent modern written language very well.

Chapter 5 ("Using parallel corpora: more advanced search procedures") moves into statistical analyses common in corpus linguistics. The chapter opens with a general discussion of whether a researcher should use statistical techniques or not, and how to go about them if one does. Various options of treating quantitative data are presented and evaluated, but the authors advocate using either desktop database software (e.g. Microsoft Access) or statistical programme packages (e.g. SPSS, R). After the generalities, most of the chapter is structured around concrete research problems and case studies exemplifying how they should be solved, as in Chapter 4. The first of these problems is checking the reliability of corpus data. The case study concerns the representation of various time periods in the literary Russian-to-Finnish part of the ParRus corpus. The second quantitative theme in Chapter 5 is measures of central tendency, to which the authors dedicate three case studies. After a quick revision of measures of central tendency, range and distribution, the section moves to case studies on sentence length in Finnish translations of Russian short stories, the dispersion of common words in the TamBiC corpus, and lexical richness in Russian novels and their Finnish translations.

The chapter then has a brief interlude on the chi-square test of independence. While often used and beginner-friendly, the appropriateness of this test in corpus linguistics has been called to question because it assumes that the observations are independent of one another, which is seldom the case in corpus data (Kilgarriff 2005; Lijffijt et al. 2016). The discussion of statistical significance testing paves the way to a more in-depth discussion of collocations. This time the definition of collocation is statistical. Two case studies are offered on the English adjective *necessary* and its translations in Finnish, one using concordances, the other so-called trans-collocations between Russian and Finnish. Trans-collocates are "collocational relationships between the aligned sentences" (p. 131). For instance, the word *bird* would have its translation equivalent as its best trans-collocate, followed by domain-specific words such as 'fly' and 'cage'.

The last technique introduced in Chapter 5 is keyword analysis, which shows what lexical items are over-represented in a given dataset when compared to a reference corpus. The case study in this section examines how well Finnish translations of Bulgakov manage to convey the author's voice. The study usefully highlights the caveats of doing this type of analysis on morphologically rich languages.

Chapter 6 ("Applications of parallel corpora") catalogues various fields of research in which parallel corpora may be of use and provides further case studies. Each section concludes with a list of sample research questions. The chapter opens with a short section on parallel corpora as dictionaries. This is followed by parallel corpora in lexicography. The case study in this section is on the Russian word *prichina* 'reason, cause' and its equivalents in Finnish. Through the example, the authors show that a very large corpus is necessary for lexicographic purposes if one wishes to go beyond the one-word level and consider the phraseologies of words. Since parallel corpora are seldom very large, the authors conclude that they cannot be the sole method for compiling a bilingual dictionary. The same applies for the topic of the following section, terminology. Here, the authors begin by introducing linguistic "laws" such as homonymy and polysemy. While potentially useful, the exposition could at times be clearer; for instance, I did not understand why *recorder* (the musical instrument) and *recorder* (an electrical appliance that records sound) are homonyms but *party* (a festive gathering of people) and *party* (political grouping) are polysemes. The case study in this section is on the terminology of the paint and varnish industry in Finnish and Russian.

The subsequent section treats morphology and syntax through the example of the Finnish present perfect translated using the English simple past. Then it is the turn of pragmatics, which is illustrated through Finnish translations of the English discourse particle *yes*. Finally, the authors exemplify translation studies by considering the sentence positions of English *however* and its Finnish equivalent *kuitenkin*. These three case studies are somewhat similar, which highlights the porousness of the boundary between the fields in question: it is not clear why the study on the translations of *yes* is a matter of pragmatics but that on *however*/*kuitenkin* an exercise in translation studies, for instance. Indeed, it might be better to conduct cross-linguistic studies of many pragmatic phenomena using comparable corpora in lieu of or in addition to parallel corpora.

Chapter 7 ("A survey of available parallel corpora") is basically a list of parallel corpora that currently exist. The chapter includes the basic

characteristics of the corpora, such as the languages involved, size, genres included and the compilers. Such a list is obviously useful, though likely to become outdated fast.[1]

The book concludes with short "final remarks", in which the authors detail their approach to writing the book as well as motivate their choice of using examples from Finnish and Russian, even though a large share of their potential audience does not know these languages. Much of this could already have been said in the preface. The final remarks are followed by a useful glossary of corpus-linguistic terms and then by two appendices, one containing a list of MA theses written at the University of Tampere and the other giving sample programmes in PHP.

The book fulfils its function as a textbook for post-graduates and beginning researchers very well. One of its virtues is that in spite of its practical orientation, it does not lose sight of the theoretical significance of parallel corpora. It is always clear why a given feature of a corpus software is worth using. The procedures are clearly explained and motivated, and there is a clear progression from basic techniques to methodologically more advanced analyses, which build on previously covered material. It is also commendable that the book consistently guides the reader to more advanced sources on the topics covered. The book is mostly well edited, although there are a few solecisms here and there that do not detract from the content, however.

On the whole, Mikhailov and Cooper have produced an introduction to parallel corpora that is clearly written and pedagogically effective. It is required reading for everyone using or compiling parallel corpora in translation studies and contrastive linguistics, but it is useful also for students and researchers in adjacent fields such as linguistic typology and applied linguistics.

References

Biber, Douglas & Conrad, Susan & Reppen, Randi. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

[1] In English corpus linguistics, an example of such a list in the form of an electronic database that is regularly updated is the Corpus Resource Database (CoRD), maintained by the Research Unit for Variation, Contacts and Change in English (VARIENG). See http://www.helsinki.fi/varieng/CoRD/.

Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276. doi:10.1515/cllt.2005.1.2.263.

Lijffijt, Jefrey & Nevalainen, Terttu & Säily, Tanja & Papapetrou, Panagiotis & Puolamäki, Kai & Mannila, Heikki. 2016. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* 31(2). 374–397. doi:10.1093/llc/fqu064.

Mauranen, Anna & Kujamäki, Pekka (eds.). 2004. *Translation universals: Do they exist?* Amsterdam: John Benjamins.

McEnery, Tony & Hardie, Andrew. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511981395.

Mortier, Liesbeth & Degand, Liesbeth. 2009. Adversative discourse markers in contrast: The need for a combined corpus approach. *International Journal of Corpus Linguistics* 14(3). 338–366. doi:10.1075/ijcl.14.3.03mor.

Teubert, Wolfgang. 1996. Comparable or parallel corpora? *International Journal of Lexicography* 9(3). 238–264. doi:10.1093/ijl/9.3.238.

Contact information:

Olli Silvennoinen
Department of Modern Language, General Linguistics
P.O. Box 24
00014 University of Helsinki
Finland
e-mail: olli(dot)silvennoinen(at)helsinki(dot)fi