Christian Jones and **Daniel Waller**. *Corpus linguistics for grammar. A guide for research*. (Routledge Corpus Linguistics Guides). London: Routledge, 2015. 202 pp. ISBN 978-0-415-74641-0. Reviewed by **Ulrike Stange**, University of Mainz.

Corpus linguistics for grammar (CLG) is a very accessible guide to corpus-based studies on English grammar. It is primarily designed for (undergraduate) students and researchers who are novices to the field of corpus linguistics and who wish to conduct their own studies on various aspects of English (lexico-) grammar. The book is divided into three parts, which consist of three chapters each. It also offers a brief introduction in which the authors outline the goals and scope of the book. After providing the theoretical foundations and presenting the working tools (Part I “Defining grammar and using corpora”), CLG introduces three different methods to analyse corpus data (Part II “Corpus linguistics for grammar: areas of investigation”) and concludes with a presentation of practical approaches to corpus-based investigations of grammar (Part III “Applications of research”). Every chapter features a number of “Try it yourself” tasks and “Sample exercises”, which are an important bonus to the book and which account for a large proportion of the volume. A section with suggested answers to all of the “Try it yourself”-exercises, a glossary listing the most important key terms and an index complete the book. The references are listed at the end of each chapter.

Chapter 1 defines what a corpus is and introduces important terminology relevant to corpus linguistics (e.g. type-token-ratio, mono-modal vs. multi-modal corpora, descriptivism). To illustrate the variety of corpora that are available to date, the authors provide a list of six corpora that differ substantially with respect to size, text types represented and date of data collection. Importantly, they state that “a small corpus can be just as effective as a large one, depending on the purpose for which it is used and the principles behind its construction” (p. 6). Jones and Waller sketch the factors that need to be taken into consideration when using (or compiling) a corpus to make sure that the data can actually provide the answers one is looking for. Using corpora to investigate actual language

use is presented as a straightforward procedure “to test and challenge our intuitions about language” (p. 9), to inform or rather improve English language teaching (ELT) by identifying highly relevant and/or frequent structures, to raise awareness of genre- or register-specific patterns, to improve dictionary entries by adding valuable information on frequency as well as collocation and colligation patterns of the individual lexical items, and to influence grammar books aiming at describing how English is used in formal and informal contexts (spoken and written). The authors stress the fact that “[t]here is a distinct grammar of spoken English” (p. 14), and that they will take this into account when discussing grammatical features in spoken and written data. Reminiscent of, but with no allusion to the approach in Construction Grammar, the authors identify the separation of lexicon and syntax as difficult. The first chapter concludes by highlighting the limits corpora have in that they provide the results but leave the interpretation of the results to the researcher, and in that, naturally, they can never be truly up to date or reflect a language in its entirety.

To begin with, Chapter 2 contrasts prescriptive and descriptive approaches to grammar on the one hand and the grammar of spoken and written language on the other. Drawing on Carter’s and McCarthy’s (2006) definition of grammar, the authors present their view on grammar, on which they gradually elaborate to include more and more relevant aspects as they progress through the chapter. This train of thought is supported by adequate sample exercises that illustrate the individual points made. Corpus-informed grammars are, of course, descriptive in nature and offer a more realistic version of how English is used by native speakers in both modalities. In this context, the chapter discusses the contrast between form and function (or rather: meaning) and introduces an adapted rank scale (based on Halliday and Matthiessen 2004) that indicates the different levels of grammar (text, sentence/utterance, clause, phrase, word, morpheme) and that, importantly, works for both spoken and written language. With reference to ideational, interpersonal and textual metafunctions (Halliday and Matthiessen 2004, 2013) the authors demonstrate that grammar is but a tool to create meaning in context and stress (again) that “[g]rammar and vocabulary are difficult to separate” (p. 29). Accordingly, they use the term *lexico-grammar* following Halliday and Matthiessen (2004) and introduce Sinclair’s (1991) *idiom principle* and *open choice principle*. The last section of this chapter discusses what kinds of information a corpus can provide and emphasises the benefits of combining qualitative and quantitative data analysis.

Chapter 3 lists selected open-access corpora (eight in total) alongside with the corresponding key information on their design and a thumbnail sketch of how they are useful for which purpose. It also describes, albeit too briefly (but

with a reference to O’Keefe *et al.* 2007 for further details), how to compile your own corpus. The authors explain the merits and functions of basic statistical tests to measure frequency effects (log-likelihood, mutual information score, t-test) and offer a short but good introduction to the free software tools LexTutor and AntConc that come in rather handy when conducting corpus-based research. The main features of these two programs are also listed for comparison to facilitate identifying their benefits and limits. Although the focus is on quantitative data analysis, Jones and Waller repeat again how important it is to also implement qualitative measures.

Chapter 4 relies heavily on sample exercises to show how the frequency of lexical items and grammatical structures can be investigated using corpus data. It draws attention to the complex issue of how to determine what counts as a token of the same type (in other words: what is a word, what are its word forms?) and how this influences the results, also with respect to the analysis of grammatical patterns. After presenting internet search engines as a convenient starting point for forming an idea of the frequency of a given grammatical structure or word, the authors highlight the benefits of using corpora instead. They also briefly discuss how many occurrences per million words (pmw) can be considered ‘frequent’ and what the different frequencies of the same grammatical pattern in different (sub)corpora can reveal about language use. Accordingly, Jones and Waller also examine the relationship between text type and frequency to identify register-specific (Biber 2012) uses. The chapter concludes with a short outline of the limits of just looking at frequency, especially with regard to ELT.

Chapter 5 is concerned with investigating chunks and colligation, again drawing on the notion that language production often relies on pre-fabricated, meaningful structures. It defines both ‘chunk’ and ‘colligation’ and shows how LexTutor and AntConc can be used to find and examine these sequences in a corpus. This chapter then discusses N-Grams, their shortcomings and how these can be remedied by also considering their colligational patterns. By contrasting political speeches with general spoken and newspaper data, it explains what insights into the register-specific priming of words and grammatical structures can be gained by analysing them in terms of chunks and colligational patterns. The chapter ends with a brief but critical discussion of the term ‘chunk’, thus offering plenty of food for thought.

Chapter 6 focuses on determining the semantic prosody of words or grammatical patterns by means of corpus analysis. After explaining the concept of semantic prosody using a variety of examples, it describes in detail how corpus data can serve to examine this feature. The authors emphasise “how central

grammatical choice is to communication” (p. 116), viz. the decision to vest a certain meaning in this or that form is context-dependent – as is illustrated by the register-specific differences pertaining to the use of personal pronouns and the passive voice.

Chapter 7 is concerned with the influence corpus linguistics has already had on ELT and with further improving it by devising corpus-informed syllabi and textbooks. It presents the benefits that accrue from consulting corpora to find out about register-specific frequencies of certain features of language, their collocational and colligational patterns, their semantic prosody as well as potential differences in the spoken and written mode. While corpus-informed learner dictionaries providing essential information on how the respective lexical items can be used in context are common these days, the number of corpus-informed textbooks is still small. Sample material from traditional textbooks serves to illustrate how their quality could be enhanced through corpus-informed description. One means of incorporating corpus linguistics in the ELT classroom could be via data-driven learning (Johns 1986, 1991) or other inductive approaches. However, these methods demand a lot of time and effort on the part of the teacher when preparing the exercises. As to first language learning, corpus-informed syllabi and textbooks could help raise awareness with respect to grammar, vocabulary and phonology because, to date, explicit language teaching to native speakers is rather restricted.

Chapter 8 introduces the reader to data-driven journalism and discourse analysis. Using political speeches as an example, it shows how corpus data can be used to examine the features ‘self-mention’ and ‘engagement marker’ in a meta-discourse framework (Hyland 2005). It describes how journalists could work with corpora when investigating certain aspects, such as how politicians exploit language as a means to achieve a certain goal (e.g. to appear more authoritative). The second part of this chapter is concerned with international discourse analysis and raises the issue of how to define ‘culture’. The frequencies pmw of *hereby* in the GloWbe corpus serve as a case in point to illustrate to what extent English-speaking countries may differ in the use of one and the same word.

Combining the methods introduced in the second part of the book, the last chapter presents three sample studies (on *if*-clauses, corpus stylistics and *bloody*) to show what potential research projects could look like using corpus data and the corpus analysis tools LExTutor and AntConc. These sample studies are prefaced by a description of the principled research process which the authors recur to when guiding the reader through the different stages in the sample studies. The authors state their research questions and formulate hypotheses,

list their results and discuss them, thus making the whole process transparent to the reader (but, of course, in a simplified manner).

Corpus linguistics for grammar is a very accessible book in terms of how it is written and how it introduces the core concepts and methodologies. It is thus perhaps more of interest to absolute beginners in the field of corpus linguistics, or to experienced corpus linguists who have a new interest in investigating grammar. Either way, basic knowledge of grammatical terminology and English grammar in general is presupposed on the part of the reader, as this volume “does not attempt to describe the main grammatical features of English” (p. 1). Rather, it offers stimulating input on how corpora can enhance our understanding of English grammar, especially with respect to grammatical features where intuition about their use and reality drift apart. It is commendable that the authors make a strong point for the differences between the grammars of spoken and written English and the concomitant need to analyse them separately, an aspect that is often forgotten or not taken heed of. Another bonus is the list of open-access corpora that includes brief descriptions of their main features and their individual advantages, as this helps the reader identify which corpus will be the most useful one to answer the research questions they have in mind. The “Try it yourself” and “Sample exercises” account for a very large proportion of the book, so that there is not much room for hard facts about corpus linguistics and grammar in general. However, this is not at all detrimental to its quality because, as the subtitle indicates, it is designed to be a guide for research, and the exercises are very helpful to put the knowledge just presented into practice. The suggested answers at the back of the book make sure the reader knows they are on the right track, and the many screenshots of the search parameters box in the *BNC* and the detailed instructions of what to put in the slots and what boxes to tick help to avoid potential misunderstandings where complex searches are concerned. The introductions to simple statistical tests and the corpus analysis tools LExTutor and AntConc are brief and to the point, allowing the reader to make efficient use of them without having to read their way through books on statistics or software manuals.

Virtually without exception, all the chapters stress the importance of including qualitative measures in the data analysis. That is, the researcher is indispensable in his/her role as interpreter of the data. Just computing frequencies and checking these for statistical significance will only provide a part of the picture. However, while this makes perfect sense, there is no need to repeat this over and over again. Instead, it would have been beneficial to show the reader what a qualitative approach looks like in detail. Burning questions such as: At what point do I need to resort to working with a sample? How do I determine what

data feed into the sample? What sample size do I need to investigate this feature? How do I actually interpret the data? What do I do with ambiguous occurrences? etc. have not been addressed. The authors should have also drawn attention to the fact that corpus data are not perfect – that is, they may contain typological errors, transcriptions of spoken language may not always be consistent, especially with regard to features that are exclusive to spoken language (for example, *Tut!*, *Tsk!* and *Tt!* all represent the dental click to express disapproval), and POS-tagging is, rather unfortunately, not absolutely reliable. Thus, even for a purely quantitative analysis a proper look at the data is essential. It would also have been desirable if the volume contained references to other open-access corpora. For instance, researchers interested in investigating grammar in the context of first language acquisition could have been pointed to the corpora available in the *CHILDES* database. The suggestions concerning corpus-informed ELT sound good, but a closer look at classroom reality reveals that their realisation is far from feasible. The authors do acknowledge this, however. While corpus-informed exercises are overshooting the mark for pupils at school, I think they are highly valuable for students of English linguistics.

In sum, despite the few points criticised, this book is highly recommendable and should be found on the bookshelf of anyone interested in investigating English grammar using corpus data.

References

- Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 3(2): 9–37.
- Carter, Ronald and Michael McCarthy. 2006. *Cambridge grammar of English*. Cambridge: Cambridge University Press.
- Halliday, Michael and Christian Matthiessen. 2013. *Halliday's Introduction to functional grammar*. London: Routledge.
- Halliday, Michael and Christian Matthiessen. 2004. *An introduction to functional grammar*. London: Routledge.
- Hyland, Ken. 2005. *Metadiscourse: Exploring interaction in writing*. London: Continuum.
- Johns, Tim. 1986. Micro-concord: A language learner's research tool. *System* 14(2): 151–162.
- Johns, Tim. 1991. Should you be persuaded: Two examples of data-driven learning. *English Language Research Journal* 4: 1–16.

- O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From corpus to classroom*. Cambridge: Cambridge University Press.
- Sinclair, John. 1991. *Corpus, concordance, collocation. Describing English language*. Oxford: Oxford University Press.