

# Review

---

Quantitative Corpus Linguistics with R: A Practical Introduction (Second edition). *Stefan Th. Gries*. London/New York: Routledge, 2017, xi + 274 pp. ISBN 978-1-138-81627-5. £115.00 (hardback).

Data processing plays an integral part in natural language processing (NLP). Particularly, with the advent of large-scale corpora and the advancement of corpus linguistics, the knowledge of data processing techniques appears to be increasingly indispensable for anyone who attempts to seek reliable empirical evidence to examine, expound, and exemplify language hypotheses and theories. In corpus linguistics, one of most challenging problems is that there is hardly any existing software tool versatile enough to tackle all the technical issues in NLP. Researchers in this area would need to learn a computer language to write user-defined programmes for their research at times. In this respect, the timely publication *Quantitative Corpus Linguistics with R: A Practical Introduction* (Second edition) is very interesting and useful because it elucidates for readers the fact that R, a robust open-source programming language and word string manipulation system, can offer language researchers a powerful and user-friendly environment to retrieve and process massive language data accurately. The intended readership of the book is broad spectrum, ranging from novice corpus linguistics researchers to experienced language engineers who are willing to add the R language to their tools for empirical analysis. Anyone who shows an interest in NLP, corpus linguistics, quantitative linguistics, computational linguistics, and data-driven general linguistics would find the book interesting. Nevertheless, it should be noted that the book is practice-based, and that the users of the book should not merely read it but should also take time to familiarize with the

operators, commands and functions before they are able to write creative scripts and programmes.

This updated volume grows out of the author's previous publication (2009). Although the topics of the books are introduced in an original way by delving into the issue of why we repeatedly introduce corpus linguistics, Gries focuses only on data retrieval, data manipulation and data evaluation, and he emphasizes that the book 'presupposes that you know what you would like to explore but gives you tools to do it that go beyond what most commonly used tools can offer' (p. 1). He overviews the contents of the books and clearly identifies the central tenets of using R programming in corpus linguistics (Chapter 1), which specifically include data processing and manipulation, text processing with/without regular expressions, fundamental aspects of statistical analysis and visualization. The overview is very functional to the extended discussion on various topics included in the book and helps construct intrinsic links between the subsequent chapters.

There are a number of advantages of using R over readily-made language data processing tools (e.g. WordSmith and AntConc) and other programming languages (e.g. Perl and Python). On the one hand, R has become a mainstream programming language, by which a large community of R users have written and posted R code and programmes for mathematical and statistical operations. This means that researchers specializing in corpus linguistics may feel interested in the code and programmes, and can use them creatively for their own research purposes. On the other hand, R is robust to perform nearly all the functions needed for language research, which are not available in other languages such as Perl and Python. Different tasks such as data retrieval, data manipulation and calculation, statistical analysis, result visualization (e.g. tables, charts, and graphs) normally require different software packages, and

now they can be simply handled once and for all with R. In addition, R is particularly useful when one attempts to employ the corpus-driven approach (see e.g. McEnery and Hardie, 2012 for a comprehensive analysis of the corpus-driven approach) to look at issues independent of predetermined grammatical items.

Addressing ways of approaching corpus linguistic studies is helpful to language researchers in this area (Chapter 2). Gries clarifies the core technical issues variously pertaining to corpus linguistic approaches. Then he discusses the methods of compiling corpora regarding corpus typology and introduces a number of basic ways in corpus analysis in terms of frequency lists, dispersion information, collocations, and concordances. The four most central methodological concepts (or tools) in corpus linguistics prompt a useful explanation as to how we should investigate lexical properties. For instance, collocation is a good way to demonstrate the ‘intimacy’ between co-occurring lexical items and can arouse language learners’ awareness of how collocating words vary in form, meaning and function (see e.g. Feng et al., 2018).

The book differs greatly from many other introductory programming books because it provides, with plain language, a down-to-earth introduction to programming syntax and programming logic (Chapter 3). This is a crucially important step for those novice programmers who wish to become more proficient in using R. Gries explains a large variety of data structures, which specifically include vectors, factors, data frames, and lists. In addition, Gries discusses elementary R functions (e.g. conditional expressions ‘else-if-{}...{}’ and ‘for-loops’) and generalizes a number of rules of R programming, which greatly help increase data processing speed and lower the risk of losing data. Apart from data structures, he emphasizes the importance of using sophisticated pattern-matching tools in character/string processing, such as accessing and changing character vectors, merging/splitting character vectors with/without regular expressions, and so forth. Compared with the first edition of the book (2009), this section is enriched with new contents, which specifically includes the discussion on Unicode and XML (extensible markup language), a

revision of the `exact.matches` functions as well as an intrinsic link with the following chapters with regard to writing creative functions in text processing.

Chapter 4 is very informative and covers a great number of topics in relation to variables involved in various corpus linguistic scenarios, such as categorical-dependent variables and numeric-dependent variables. Apart from the manageable introduction of basic R knowledge, Gries shows readers the methods of carrying out statistical analysis using R functions and visualizing empirical findings with self-evident tables, charts and graphs. All of these merits would serve as an ideal reference for researchers to design new projects and make the methods mentioned in the book well situated within any instruction involving corpus linguistics.

Drawing attention to the application of R to different central tasks in NLP, Chapter 5 appears to be the most important section of the book. It shows how R functions can be applied to the investigation of commonly discussed topics in corpus linguistics, such as dispersion, retrieval of N-grams, collocation and colligation, concordance, type-token ratio, vocabulary growth, and so forth. Adopting more than 30 case studies, Gries suggests that the tasks performed can be grouped in a heuristic way, and that R users would need to be familiar with the functions they have already accessed and combine them into complicated scripts. Readers are frequently required to consider how they specifically manage language data by creating arrays and hash tables, combine R functions and write creative programmes. Such heuristic grouping is very useful for readers to follow the step-by-step instructions provided, in which R code is tabulated for simple and clear explanations of each argument in the code. The final chapter wraps up the book by summarizing the potentials of R programming in quantitative corpus studies. Gries suggests readers check a few relevant packages that may help with NLP, such as `stringi`, `stringer`, `gsubfn`, `stringdist`, `tm`, `openNLP`, and `koRpus`. The section of appendix offers additional resources for corpus analysis, which includes companion websites (both corpus-linguistics and statistics with R), R and R Add-ons, available software, and regular expression testers and websites. This indicates that readers can communicate with

other R users online and can obtain timely advice from the author when they meet technical difficulties in R programming. The exercises provided on these websites cover a wide spectrum of topics concerning language studies, such as morphology, syntax, semantics, and pragmatics, to name but a few.

A number of major strengths of the volume should be noted. First of all, the methodological innovations demonstrated in the book have breathed new life into the traditional approaches to corpus linguistics. Although Gries repeatedly emphasizes that this book is only designed for teaching readers to process language data with R, the way he organizes R programming techniques is well situated within the introductory framework of corpus linguistics from both theoretical and practical perspectives. Secondly, the book allows for both fundamentals and potentials in NLP. There may be some issues common to all because the volume presents a number of case studies involving wordlist, frequency, collocation, colligation, concordance, *N*-grams as well as statistical measures. Nevertheless, Gries helps readers realize that, apart from the essentials in NLP, they can tackle a wider range of tasks using R, such as vocabulary richness, vocabulary growth, verb valency, dependency distance, polysemy, and so forth. Thirdly, given computer programming is still an emerging area in relation to corpus linguistics, readers may be inspired by the novelty of many aspects of the volume. With respect to this, follow-up exercises and studies may bring forth interesting results by using the novel techniques or approaches to other languages or different texts. A total number of more than 200 references at the end of each chapter would greatly help readers in various directions where the R language can be applied in NLP.

Compared with the precedent edition (2009), this updated volume is restructured in a more logical order and includes a large number of topics pertaining to corpus studies involving R programming, such as dispersion. Compared with other similar publications such as *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics* (Desagulier, 2017), the book under review may appear to be more suitable for those who have acquired elementary programming knowledge or

who have had previous experience in using other computer languages. With respect to readability, the book features precise and concise academic language, in which Gries gives readers ‘think & break’ time regularly to consider what they can do with the R functions. All of these merits make this practice-based guidebook very user-friendly because readers would feel like sitting *vis-à-vis* with the author when reading the book. From this perspective, a large group of audience can benefit from the book, particularly those specializing in NLP, corpus linguistics, quantitative linguistics, and computational linguistics. Therefore, the book remains a great contribution to the area of NLP and deserves my wholehearted recommendation.

## Acknowledgements

This work was supported by Confucius Institute Headquarters (Hanban)/Chinese Society of Academic Degrees and Graduate Education and National MTCOSOL Education Steering Committee [HGJ201706]; Liaoning Office for Education Sciences Planning [JG18DB007]; Liaoning Provincial Federation Social Science Circles [2019slsktyb-052]; and Postgraduate Office of Bohai University [02200104441-44].

## References

- Desagulier, G. (2017). *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Cham: Springer.
- Feng, H., Crezee, I., and Grant, L. (2018). Form and Meaning in Collocations: A Corpus-driven Study on Translation Universals in Chinese-to-English Business Translation. *Perspectives*, 26(5): 677–90.
- Gries, S. T. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. London/New York: Routledge.
- McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Haoda Feng   
School of Foreign Languages, Bohai University, P. R. China  
doi:10.1093/llc/fqz021