REVIEW: Gries (2009) *Quantitative Corpus Linguistics with R*. London and New York: Routledge

Victoria Clark-Sánchez [1]

There are many teachers and scholars who value corpus analysis, but do not yet do their own programming for teaching and research. Many have been considering making the leap from using only pre-packaged concordancing software to writing our own programs, and Gries's *Quantitative Corpus Linguistics with R* may be the resource that helps many of us to make the transition.

It may be difficult, however, to see the advantages of doing one's own programming until one has conducted corpus research using only concordancers and developed a sense of their limitations. First, I will allow that concordancers are very useful when the focus of the research is on very specific, pre-determined words or phrases, or other top-down sorts of research questions. Concordancers are a good way to introduce students to the concept of corpus-informed language learning and they demonstrate the way that near-synonymous lexical items vary in their functions. These uses have been garnering increasing attention in the wider fields of Applied Linguistics and even Theoretical Linguistics, and we can thank concordancing software for making corpora more accessible and opening up the tool of corpus analysis to an increasing number of students, teachers and researchers.

However, when one starts to ask deeper questions about patterns of language use, and we desire a more bottom-up approach to linguistic analysis that is not dependent on pre-determined lexical or grammatical items, the limitations of concordancers quickly become apparent. To limit our linguistic inquiry only to questions that can be investigated through pre-packaged software is akin to limiting our study of syntax to items that can be identified by the green squiggly line of the grammar checker in our word processor. We are at the mercy of the software developers in terms of the analytical features included and the price to purchase or upgrade these tools.

[1] Department of Communication Sciences and Disorders, Northern Arizona University, PO Box 15045, Flagstaff, AZ 86011, USA
Correspondence to: Victoria Clark-Sánchez, *e-mail*: vec@nau.edu

Corpora 2013 Vol. 8 (2): 269–272
DOI: 10.3366/cor.2013.0045

www.euppublishing.com/cor

The typical reader who picks up Gries's new book would probably fall into one of the following groups:

(*1*) A Linguistics and/or Applied Linguistics student or scholar who is new to corpus linguistics and to programming (beyond the scope of available concordancing packages);
(*2*) A corpus linguist who is considering adding R to his/her tools for analysis and is already familiar with programming; and,
(*3*) A computational linguist who is new to corpus linguistics methodology and goals.

For each of these groups, different accommodations must be made, and Gries makes a fine job of keeping his audience engaged. While this book has a lot to offer to all of these groups, it is perhaps best suited to groups (2) and (3) for the purposes of self-study. For those in group (1), with little or no programming background, the material would be easier to study in a classroom, study group or with a mentor. For those with some programming experience, an essential question must be answered, which is: Why R and not Perl or Python or some other language that many are more familiar with?

Perl and Python are similar to R in that they are free and open-source, which means that users all over the world can post free, ready-made programs or bits of code that perform functions that are of interest to linguists. There are programmers using all three languages for research into linguistics, and the individual user, especially the beginner, is best served by using the language (Perl, Python, R or a host of others) for which mentoring is most readily available, since all of these languages are capable of processing language for linguistics research. The main advantage of R is that it has become the programming language of choice in mathematics and so there is a large community of users who have created and posted code for advanced mathematical and statistical operations that may be of interest to linguists. In addition, it is well-suited to performing all of the functions that are needed for research from beginning to end: extracting data from the corpus, performing calculations and statistical analysis, and constructing graphs and charts that are ready for publication. To do all of these functions without such a tool, we currently need to use several software packages, including a word processor, a concordancer, a spreadsheet program and a statistical program. In the introduction of the book, the author provides additional detail in comparing R to its alternatives.

Gries designed this book with his Intro to Corpus Linguistics students in mind, and the material on corpus linguistics as a tool for linguistic analysis in Chapters 1 and 2 provides a very brief summary of the field, including descriptions of types of corpora, the use of frequency lists, collocations and lexico-grammatical co-occurrence.

In Chapters 3 and 4, Gries gives the reader everything needed to get started. The software is free and there is an active community of R developers who process language and have made useful code available. Processing texts

that are written in non-English non-ASCII characters is covered here, using the example of Russian/Cyrillic. This book stands out from many other introductory programming books in that it introduces programming syntax, as well as the building blocks of programming logic, or how to look at language through the eyes of a computer. This is essential for those who are new to programming and is certainly the most uncomfortable part of becoming proficient in one's first computer language. Gries does a fine job with this although the density of programming concepts and lingo in these two chapters is best tackled with a more experienced programmer (teacher, tutor or mentor) by all but the most intrepid beginners.

Chapter 5 introduces one of the most attractive features of R, which is the ability to run statistical analysis and generate charts, tables and graphs of the results. Gries provides a concise introduction concerning the goals of quantitative analysis, hypothesis formation, research design and choosing statistical analysis, which would be complimented well in an Intro to Corpus Linguistics class by Biber *et al*. (1998) and McEnery *et al*. (2006). Interpretation of the output and presenting output visually are also covered. This chapter functions as a manageable introduction of the essentials of statistical analysis for corpus linguistics, and can serve as a quick-reference for experienced corpus linguists designing new studies.

In Chapter 6, the author points the reader to case studies that can be accessed online on a companion website to the book and provide further practice in the techniques presented therein. He also maintains a companion website and newsgroup for the book with additional exercises and a way for readers to communicate with one another and with the author. The sample exercises on the website cover many areas of linguistic inquiry such as morphology, syntax, semantics/lexicography, pragmatics/text linguistics and several other sub-categories.

The Appendix immediately following Chapter 6 lists over sixty additional resources for corpus analysis, including freely available software, indexes of regular expressions, available corpora and journals that publish corpus research.

The methodology of corpus linguistics is an intersection of scholars from many different fields, and is used for multiple purposes within each. There are some issues common to all, however, and Gries helps readers 'get their feet wet' with the essentials of corpus analysis. It is easy to imagine a group of students finishing this book over a semester and being able to complete their own corpus linguistics studies by the end of the semester. About 200 references at the end of the book give the reader a good start in many directions where corpus methodology is applied.

The subject index of this book is useful for finding the topics and authors covered in Gries's book. However, a glossary of terms, or a table of frequently used functions, would have been useful for the user encountering new terms for the first time, or would allow the user to dip in for a quick reminder of the syntax of a function like 'strsplit', for example – though most users may be just as comfortable searching for such information online.

Gries has completed a new book that serves as a logical follow-up to this volume. This book is an expansion of the statistical analyses used in corpus linguistics and examples of how to conduct these analyses using R, including multivariate statistics and regression.

References

Biber, D., S. Conrad and R. Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

McEnery, T., R. Xiao and Y. Tono. 2006. Corpus-based Language Studies: An Advanced Resource Book. London and New York: Routledge.