BOOK REVIEW

Doing quantitative corpus linguistics R-ight.

Gries, S. Th. 2009. *Quantitative Corpus Linguistics with R. A Practical Introduction*. New York/London: Routledge. (vii + 248 pp.)

In a recent study of his (Gries 2010: 123), the author of the book under review poses two questions:

(1) Why is it that we corpus linguists look at something (language) that is completely based on distributional/frequency-based probabilistic data and just as complex as what psychologists, psycholinguists, cognitive scientists, sociolinguists, etc. look at, but most of our curricula do not contain a single course on statistical methods (while psychologists etc. regularly have two to three basic and one or two advanced courses on such methods)?

(2) And why is it that we as corpus linguists often must retrieve complex patterns from gigabytes of messy data in various encodings and forms of organization, but most of our curricula do not contain even a single course on basic programming skills or relational databases (while psychologists, computational linguists, cognitive scientists etc. devote years to acquiring the required methodological skills)?

The questions are of course rhetorical questions. Rather than demanding an answer they point to what the author sees as the lamentable status quo in (much of) corpus linguistic practice. The aim of the book under review is to contribute to remedying the situation. The contribution offered in response to question (1) lies in providing an introduction to inferential statistics; the contribution offered in response to question (2) lies in providing an introduction to R, a programming language and statistical environment for doing inferential statistics.

The author pursues these aims within a broader conception of the book as "an introduction to corpus linguistics" (p. 1). Indeed, following the Introduction, which states the book's aims, outlines its contents, and gives recommendations for instructors, Chapter 2 defines the notion of a corpus and briefly presents what the author takes to be the three central corpus-linguistic methods, namely "frequency lists, collocations, and concordances" (p. 4). Given that Chapter 2 is only 11 pages long, the book, as an introduction to corpus linguistics, will not be a serious

International Journal of Corpus Linguistics 17:2 (2012), **287–293**. DOI 10.1075/ijcl.17.2.06ruh
ISSN 1384–6655 / E-ISSN 1569–9811

competitor for other introductions to corpus linguistics which deal with matters relevant to corpus linguistics in much more detail. The brevity and sketchiness of the chapter follows the author's intention to save space for his introduction to R and is partly made up for by the inclusion at the end of the chapter of a "further reading" box containing useful references to more in-depth publications on fundamental aspects of corpus linguistic research (a strategy that is consistently practiced throughout the book).

Crucially, it is worth considering Gries's dictum made in that first chapter that "strictly speaking at least, the only thing corpora can provide is information on frequencies"(p. 11) […] "and it is up to the researcher to interpret frequencies of occurrence and co-occurrence in meaningful or functional terms" (ibid.).[1] If one accepts this (undoubtedly provocative) premise, one will also have to accept its entailments: that "strictly speaking at least", (i) corpus linguistics is an essentially quantitative science, and that (ii) corpus linguistics, as other quantitative sciences, should adopt methods to evaluate quantitative findings statistically, that is, with a view to their generalizability from the limited sample (the corpus) to the unlimited population (the language). The remainder of the book, Chapters 3 to 6, homes in on one tool that can be usefully exploited for doing corpus linguistics in the intended quantitative way, the programming language R which was developed in the 1990s at the University of Auckland, New Zealand, and which, as the author emphasizes, allows you "to perform just about all corpus-linguistic tasks within only one programming environment" (p. 3) including "data processing, data retrieval, annotation, statistical evaluation, graphical representation …" (p. 3). Chapter 3 gives the reader instructions as to how to download and install R. It also introduces the most fundamental functions and commands in R. Further, it deals in great detail with central data structures such as vectors, data frames, and lists, and describes how to load and access (parts of) them. Moreover, the chapter turns to some elementary programming functions available in R. Finally, it explores R's pattern matching tools for the processing of character strings. Chapter 4 promises to teach the reader "how the functions introduced in Chapter 3 can be applied to central tasks of a corpus linguist, namely to retrieve and process linguistic data to generate frequency lists, concordances, and collocate displays" (p. 105). Chapter 4 and Chapter 3 are technically quite demanding and will be hard to follow for readers not yet sufficiently familiar with R. Underlying the length and depth at which these technical details are explained is arguably the author's wish, related to question (2) above, to prove that R can serve as "the corpus linguist's all-purpose tool" (p. 2), that is, as the only, or at least major, tool he or she will ever need for any corpus linguistic task at hand.

Chapter 5 is intimately related to the author's second central point related to question (1) above (that corpus linguistics, as a quantitative science, must embrace

statistical methods). It introduces the reader in great care and with compelling clarity to statistical thinking, the nature of different types of variables (categorical, binary, continuous, and ordinal) as well as their type of relation to one another (dependent or independent). It also defines standards in formulating and operationalizing hypotheses and reporting results. The fundamentals of statistical thinking discussed by the author include the 'falsification paradigm' (an alternative hypothesis cannot be positively proven; it can only be preferred to its logical counterpart, the null hypothesis, if the probability that the latter describes the data adequately turns out to be vanishingly low — less than 5 per cent). Another fundamental notion is the necessity to assess the significance of findings via a juxtaposition of the observed values against the expected values (for a poignant discussion of this issue, see Stefanowitsch 2005). The chapter also discusses some basic statistical tests, including the chi-squared test, the Shapiro-Wilk test for normality of distribution, the Ansari-Bradley test for homogeneity of variances, the Wilcoxon rank sum test, and the Kendall correlation test. Special attention is paid to the chi-squared test most commonly reported in corpus studies: the author exemplifies its two major variants — the chi-squared test for given or equal probabilities (one categorical variable with two levels; for example, frequencies of the article *the* in two same-size samples) and the chi-squared test for independence (one dependent and one independent categorical variable; for example, frequencies of *the* in relation to whether men or women use it). Chapter 5 also demonstrates R's huge potential for visual representation of data and distributions. For example, the author devotes some worthwhile time on the boxplot, an extremely useful graphic for summarizing a great deal of information on the location and dispersion of values of categorical variables. This information includes: the median (the "middle" value of the sorted values of a distribution), the interquartile range (IQR), that is, the 50 per cent of the data grouped around the median, 1.5 times the IQR (shown in the "whiskers", the dashed vertical lines with the horizontal limits), outliers (data points considered atypical given their unusual distance from the bulk of the data), and optionally, a confidence range shown in "notches" (the range within which the "true" median is likely to occur). For illustration, consider the lengths of utterances (as measured in number of words) in the textual components of the Narrative Corpus (NC), a corpus of conversational stories (cf. Rühlemann & O'Donnell forthcoming) depicted in the boxplots in Figure 1. Textual components in the NC consist of the following: pre-narrative conversation (CPR) (turn-by-turn talk preceding the story), narrative-initial utterance (CNI) (the utterance which launches the story), narrative-final utterance (CNF) (the utterance immediately preceding the participants' re-orientation to the present speech situation), narrative-medial utterances (CNM) (all utterances between CNI and CNF), and post-narrative conversation (turn-by turn talk following the story).

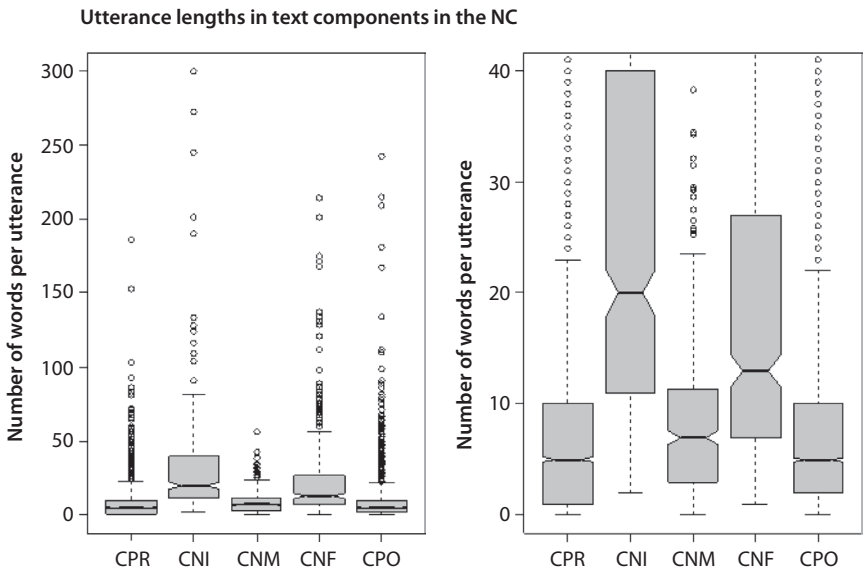

Figure 1. Boxplots of lengths of utterances across micro-components

While the left panel in Figure 1 shows the distributions with all outliers (indicated by empty circles) included, the right panel presents the boxes in greater resolution (note the altered scale on the y-axis). In the left panel, it can be seen that the dispersion of the lenghts of narrative-initial utterances (CNI) and narrative-final utterances (CNF) is much greater than in the pre- and post-narrative components (CPR and CPO) as well as in narrative-medial utterances (CNM). This greater dispersion, and hence greater heterogeneity, is suggested not only by the larger range, for CNI, of outliers, but also, for CNI and CNF, by the greater distances of the whiskers and the larger sizes of the boxes (indicating wider IQRs). Also, as shown in the right panel, the medians (indicated by bold horizontal lines cutting across the boxes) for CNI and CNF are much higher than for CPR and CPO, suggesting that narrators take longer turns both when setting (CNI) and closing (CNF) the scene of the story than conversationalists do in non-narrative conversation. Further, only the notches (indicating the confidence intervals) for CPR and CPO overlap, but not those for the three narrative components. This is visual evidence suggesting that the differences in medians are significant. This is confirmed by pairwise comparisons using Wilcoxon rank sum test, where all but one of the pairings (namely CPR and CPO) exhibit significantly different distributions. Turn size is therefore clearly a function of genre (general conversation vs. conversational narrative) and position within that genre (narrative-medial vs. narrative-initial/-final) (for a more detailed discussion, see Rühlemann forthcoming). The usefulness of the boxplot, once one has learned to "read" it, lies in its capacity to

reveal all this crucial information on distribution, dispersion, and potential significance at a single glance.

Other useful graphical representation types shown in the book include: interaction plot (pp. 179–181), bar plot (pp. 186–187), mosaic plot (p. 194), association plot (p. 198), line plot (p. 202), and scatter plot (p. 212). To me, this chapter is the heart of the book; it will make or break the author's argument: that doing statistics is necessary and worth it and that it's by using R that quantitative corpus linguistics is done R-ight. The concluding chapter, Chapter 6, very briefly rounds off the book. It introduces the reader to case studies available at the companion website that aim "at bringing together various things explained and exemplified in previous chapters" (p. 219). Chapter 6 also considers further corpus-linguistic applications of R in related fields such as text linguistics, pragmatics, and applied linguistics.

The book is effectively structured as a practical introduction. It is made clear from the start that the many methodological and technical skills the book aims to teach the reader cannot be acquired by reading it alone. Rather, the book, like other comparable introductions centrally concerned with teaching technological skills — for example, Hoffmann et al.'s (2008) introduction to corpus linguistics with the BNCweb interface — is conceptualized as a manual in the sense that "you must read this book while sitting at your computer and directly entering the code and working with the code" (p. 5). In line with this design as a how-to instruction, the book contains features intended to increase the reader's interactivity. These features include the numerous passages, distinguished from the text by gray shading and different font type, that contain nothing but the R code, so-called "think breaks" which the author introduces with concise and thought-provoking questions, as well as exercise boxes with small assignments.

Further, a remarkable characteristic of the book is its unusual type of intertextuality. Unlike other textbooks that are self-contained in the sense that all the information needed to benefit from them is (or should be) in the book itself, this introduction is at the centre of an extended network of related "texts". These include a companion website from where the readers are encouraged to download the data files for the worked examples as well as the exercise assignments (http://www.linguistics.ucsb.edu/faculty/stgries/research/qclwr/qclwr.html) and a Google Group called "CorpLing with R"(http://groups.google.com/group/corpling-with-r) the author created and maintains. More loosely connected to the network is the author's second textbook on R (Gries 2009), which is targeted to a wider linguistic audience and much more comprehensive in the sense that it discusses a much broader variety of R functionalities, including most notably how to do multifactorial studies in R, a topic not addressed in the present book. Finally, situated in the wider periphery of the network, the author's "bootcamps" related to R should be mentioned, which have inspired not only a fierce debate on how corpus linguistics

defines itself (cf. Worlock Pope's 2010 Special Issue of this journal *The Bootcamp Discourse and Beyond*) but also (maybe more importantly) introduced hundreds of linguists around the world to the delights of R.

In sum, it would be wrong to suggest that this book is a must for every corpus linguist's bookshelf for, being a decidedly practical introduction to R, it is not on the shelf that it fulfils its purpose. More correctly, this book is a must for every corpus linguist's desk right beside the computer — that's where it does its wonders. Whether or not the author will succeed in convincing his readers that R is so versatile a tool that they will no longer have to use any other tool remains to be seen. The most significant gain readers can make from this book is being shown around in the universe of inferential statistics and sophisticated graphical representation. This universe is not only exciting, offering the desperately needed look beyond the confines of the data into the population (language), but it is also the natural habitat of corpus linguistics, which is, as noted, largely quantitative in nature. It is high time the discipline explored this universe in much greater depth.

Note

1. Note that the dictum is "prefaced" by "strictly speaking" — a hint that the notion is not to be taken dogmatically. Thus, the author tacitly, as it were, acknowledges that corpora can also provide other things than frequencies; see, for example, the largely qualitative research on semantic prosody (for an initial attempt at quantifying "good" and "bad" prosodies, see Dilts & Newman 2006).

References

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund Prytz, Y. 2008. *Corpus Linguistics with BNCweb — A Practical Guide*. Frankfurt am Main: Peter Lang.

Dilts, P. & Newman, J. 2006. "A note on quantifying 'good' and 'bad' prosodies". *Corpus Linguistics and Linguistic Theory*, 2 (2), 233–242.

Gries, S. Th. 2009. *Statistics for Linguistics with R. A Practical Introduction*. Berlin: Mouton de Gruyter.

Gries, S. Th. 2010. "Methodological skills in corpus linguistics: A polemic and some pointers towards quantitative methods". In T. Harris & M. Moreno Jaén (Eds.), *Corpus Linguistics in Language Teaching*. Frankfurt am Main: Peter Lang, 121–146.

Rühlemann, C. Forthcoming. *Narrative in English Conversation: A Corpus Analysis of Storytelling as an Interactional Achievement*. Cambridge: Cambridge University Press.

Rühlemann, C. & O'Donnell, M. B. Forthcoming. "Introducing a corpus of conversational narrative. Construction and annotation of the *Narrative Corpus*". *Corpus Linguistics and Linguistic Theory*.

Stefanowitsch, A. 2005."New York, Dayton (Ohio), and the raw frequency fallacy". *Corpus Linguistics and Linguistic Theory*, 1 (2), 295–301.

Worlock Pope, C. (Ed.) 2010. *The Bootcamp Discourse and Beyond*. Special Issue of the *International Journal of Corpus Linguistics,* 15 (3).

Reviewed by Christoph Rühlemann,
Ludwig-Maximilians-Universität, Munich