RiCL Research in Corpus Linguistics

Review of Wallis, Sean. 2020. *Statistics in Corpus Linguistics: A New Approach*. London: Routledge. ISBN: 978-1-138-58938-4. DOI: https://doi.org/10.4324/9780429491696

Tove Larsson
Northern Arizona University / United States

In this book, Sean Wallis provides an introduction to statistics as it applies to corpus linguistics studies. The book offers a valuable complement to existing resources available to researchers in the field, (i) by focusing on distributions and confidence intervals and (ii) by offering in-depth explanations of their underlying mathematics and logic. As stated in the preface, "what seems missing [in traditional books on statistics for corpus linguistics] is a clear explanation as to how a test procedure works from ground up" (p. xiii). The book has 19 chapters divided into six sections: 1. "Motivation," 2. "Designing experiments with corpora," 3. "Confidence intervals and significance tests," 4. "Effect sizes and meta-tests," 5. "Statistical solutions for corpus samples," and 6. "Concluding remarks." This review will follow the same outline and end with a brief evaluation of the volume.

In the first section, Wallis lays the foundations for the subsequent chapters by providing an overview of the field of corpus linguistics, different kinds of corpora, and study designs that researchers who work with corpora may use. He introduces three distinct classes of empirical evidence that can be obtained from a corpus: factual evidence of a linguistic event (i.e., a linguistic token), frequency of a linguistic event, and interaction evidence between two of more linguistic events (i.e., the probability that an event *x* will occur, given an event *y*). After introducing what is referred to as the '3A Cycle' (Annotation, Abstraction, and Analysis), which is central to all corpus linguistic studies, Wallis subsequently addresses the question of what (richly) annotated corpora can tell us. The author also explicitly argues against a "simplistic 'bigger is best'

DOI 10.32714/ricl.10.01.13

approach" (p. 3) to data analysis and corpus building. The section concludes with some example studies and a discussion of framing constraints in study design.

The second section discusses how the scientific method is applied in corpus linguistic studies. It begins with an overview of the research process, including the compilation of a corpus (i.e., a sample), formulation of research questions and hypotheses, and evaluating these hypotheses through experiments and statistical tests. In the process, the author introduces concepts such as variables (and the fact that the variables we use are of different numerical types or scales: binomial, multinomial, ordinal, interval, and real). The author then widens the discussion to variationist designs, that is, designs where linguists study "the influence on [decisions that speakers or writers make when forming utterances], identifying factors that affect the selection of one option over another" (p. 47). Binomial and multinomial techniques are introduced and illustrated using linguistic examples. The author argues that such designs are preferable to designs that rely on word-based baselines (e.g. per-million-word frequencies), as "language is not a sequence of random words" (p. 48) and as such baselines do not "distinguish opportunity and choice, and are vulnerable to arbitrary variation" (p. 74). Finally, sampling as it applies to corpus linguistics is discussed. Example studies are used to illustrate the techniques and points made throughout.

Section 3 is devoted to inferential statistics with a specific focus on distributions and confidence intervals. Wallis builds on the discussion in Sections 2 and 3 to introduce foundational concepts such as *p*-values, distributions, and confidence intervals. Using the Wilson score interval and the Newcombe-Wilson Interval, the notion of 'significant differences' (in the null hypothesis statistical testing framework) is introduced. This section also includes a chapter on replication and 'the replication crisis' (which started with the observation that findings from many studies are difficult to reproduce). The author brings up possible reasons why findings in corpus linguistics studies may not replicate, such as differences across studies with regard to populations, samples, and/or operationalizations of key categories and constructs. He further gives recommendations for the field moving forward, including a checklist for empirical linguistics studies that emphasizes the need for accuracy and transparency in our reporting practices. The final chapter in this section deals with the question of how to choose the right statistical test depending on the type and scale of the variables of interest.

In Section 4, Wallis covers a discussion of effect size measures and meta-tests. First, measures of interdependence (i.e., measures of the bidirectional association between independent and dependent variables) such as Cramér's Φ are discussed. The author then moves on to introduce tests that can be used to compare results across studies: so-called 'meta-tests'. As pointed out by the author, such tests are helpful in the context of replication in that researchers may wish to compare results from (a) studies for which the data are kept constant, but where the design has been changed slightly, or (b) studies for which the design is kept constant, but where the data are different. Specifically, different Wilson-based tests are introduced and exemplified using linguistic data.

Section 5 discusses different statistical solutions for corpus samples. Wallis begins this section by stressing the importance of being able to justify the frequencies obtained from a corpus analysis. That is, we should not take output from taggers and concordancers at face value, but rather always assess and try to improve the accuracy of the annotation of the phenomenon of interest, especially for larger corpora that have not been manually checked. The author goes on to propose a golden rule of data: "We need to know that, as far as possible, our dataset is a sound and complete set of examples of the linguistic phenomenon in which we are interested" (p. 263). The remainder of the section is devoted to a discussion of how to recalibrate tests such as binomial models to account for sampling issues common to corpus linguistics (such as the fact that in many cases, the assumption of case independence is violated).

The sixth and final section of the book contains two chapters. The first includes an in-depth description of Wilson distributions (previously discussed in the volume in the context of the Wilson score interval) with the purpose of having the reader "understand the performance of the Wilson formula, distribution and interval itself" (p. 297). The impact of the size of the sample is also discussed. The final chapter of the book offers concluding remarks where the author comments on the content of the book and what he would like to see for the field moving forward. For example, the importance of making sure we have a reliable sample/data source is stressed: "if our source data are not what we think they are, all statistical generalisation must be in vain!" (p. 315).

From an evaluative perspective, this volume has a number of considerable strengths. First and foremost is the fact that the book is written specifically for a corpus linguistic audience using example studies and data from the field. As anyone who has taken statistics courses in other field knows, learning about a new technique using data and study designs that are unfamiliar adds a layer of difficulty to a subject that can already be challenging, requiring the learner to 'translate' the new techniques between fields and figure out how they apply to our kinds of research questions. It is thus immensely helpful to have a resource such as the present volume that is written specifically with our kinds of data and analyses in mind.

Another clear strength of the volume, one in line with its stated goal of making inferential statistics more tangible to help readers understand what we are doing at all stages of the analysis, is its detailed illustrations of techniques and processes. The graphs and sample studies that are spread out across the sections are very helpful for making the statistical reasoning more concrete. In this context, Wallis makes important methodological points about the importance of rigorous study design (including sampling, corpus annotation, and analysis), as "statistical methods do not turn a poor experimental design into a good one" (p. 316). He even goes so far as to say that "significance testing is secondary to the primary task of plotting data and engaging with a linguistic evaluation of what our results might mean" (pp. 314–315). Because statistics, by some, may be perceived as necessarily involving an element of 'black-box-iness', this is a refreshing perspective, one where the main focus is on the linguistic analysis and where the statistical analysis is viewed as a tool that enables generalizations of the results of that analysis beyond a specific sample.

The subtitle of the book, *A New Approach*, is indeed apt in that, unlike previous books that focus more on traditional significance testing techniques, readers of this book get to view statistics primarily through the lens of distributions and confidence intervals. While this approach has many advantages (some of which are outlined above), one downside is that the book does not cover many of the techniques that researchers in the field would encounter in publications and courses and that they may therefore benefit from learning more about. To be fair, as stated in the preface, Wallis's book is intended to "supplement, not replace other textbooks in statistics or linguistics" (p. xiv). As such, the ideal audience for the book may thus not be complete beginners in either corpus linguistics or statistics, but rather, perhaps, readers who already have some foundation in both and who wish to improve and complement their statistical reasoning and understanding in the context of corpus linguistics. Regardless of their level, however, it would perhaps be helpful for readers if future editions of the book could include some additional instructional materials (maybe as online supplements), such as exercises and code for one or several software packages, to help readers get started applying their newfound knowledge.

Further, and related to the previous point, one of the main strengths of the books is, perhaps, also a limitation: the book covers a lot of ground in relatively few pages (approximately 350 pages). That is, while organized in a fairly logical manner, the book spans an impressive set of topics, both corpus linguistic and statistical, which inevitably means that some topics must be covered in more detail than others. At times, this requires readers to have fairly advanced knowledge of topics to be able to follow the line of argumentation. Some more sign-posting, interim summaries, and suggestions for further reading would have been helpful here.

All in all, Wallis offers a very interesting and —to corpus linguistics— new perspective on statistical analysis. In addition, the many points the author makes throughout the volume on the importance of transparency and rigor in our study designs are all well taken and tremendously important for a field that finds itself growing —and needing to grow— quantitatively. There is no doubt that this volume constitutes a very valuable resource for current students and researchers in corpus linguistics, one that will no doubt also continue to grow in future editions to meet our field’s exciting future.

Reviewed by
Tove Larsson
Northern Arizona University
English Department
Box 6032
Flagstaff, AZ 86001
United States
e-mail: Tove.Larsson@nau.edu