Natural Language Engineering (2023), **29**, pp. 177–180
doi:10.1017/S1351324921000413

CAMBRIDGE
UNIVERSITY PRESS

Book Review

Statistics in Corpus Linguistics: A New Approach by Sean Wallis. New York/Oxon: Routledge, 2021. ISBN 9781138589384 (PB: 44.95), ISBN 9781138589377 (HB: 160.00), ISBN 9780429491696 (eBook: 44.95), xxvi+382 pages.

As the title suggests, this book under review occupies a unique niche among many other volumes discussing the use of statistics in corpus linguistics because of the "new approach" it proposes. Instead of a simple and long-formed theory introduction, each chapter contains a question accompanied by authentic research cases for further explanation. The entirety of the book is well worth reading, from the preface to the list of references. Professor Sean Wallis, the author of the book and a veteran who has over 25 years of experience in linguistics and data science, believes that "short cuts can be risky (p. 27)". To that end, although this book does not cover the full range of possibly relevant guidance to statistics and data processing, it advocates the systematic *methodology* learning rather than a simple duplication of various *methods*.

Apart from the preface, glossary, reference list, and index, this book consists of 6 parts, comprising 19 chapters. It is worth mentioning that the glossary section makes the book appropriate for beginners. Some data-phobic novices in statistics may find the book comforting enough with many professional concepts explained in the glossary for reference. Part 1 sets the scene for the motivation and explains the reasoning behind building corpora and extracting useful data from them. Part 2 demonstrates the crucial steps of designing the experiments with corpora, accompanied by the corresponding underlying logics. Parts 3 and 4 serve as the main course of the feast – the introduction to linguistic statistics knowledge, from basic to in-depth. Part 5 is devoted to statistical solutions for some corpus examples. Part 6 contains two sub-sections, one attempting to plot the Wilson distribution, whose interval is inverted based on the famous bell curve Normal distribution; and the other presenting the concluding remarks of the whole book. All parts are written in a clear-cut and vivid style, catering to its markedly diverse foci and target readership.

The first, background, part primes readers with the basic viewpoints of the author for corpus statistics. There are three noteworthy points about Part 1. At the very beginning, the author emphasises his tenet of corpus study, that is, bigger is definitely not best. In his opinion, what a linguist truly needs is a trade-off between the size of the corpus and its annotation richness. It has been proved that different annotation schemes have an impact on the knowledge that a corpus linguist can retrieve (Yan and Liu in press). Therefore, an annotation framework is fundamental for building a corpus, and a language researcher must critically engage with any given annotation scheme. Second, the author argues against some generative viewpoints that corpus research is doomed to investigate surface phenomena (e.g. Aarts 2001). He directs the reader towards the main contention of the book, that corpora should play a role in assisting psycholinguistics through the complementary source of evidence they offer. Last but not least, the author stresses the necessity of building and working with spoken corpora. Confirming these principles, two simple corpus linguistics experimental designs, namely *The British Component of the International Corpus of English* (*ICE-GB)* and the *Diachronic Corpus of Present-day Spoken English (DCPSE),* are presented for detailed illustration. Before that, the author also depicts the usage-based view of the typical 3A perspective (*Annotation*, *Abstraction* and *Analysis*) in corpus linguistics (Wallis and Nelson 2001). Additionally, *Fuzzy Tree Fragments (FTF)*, a tree-shaped diagrammatic query system, is presented via a set of screenshots demonstrating the processing steps with relevant platform tools. With this introduction to the 3A paradigm and *FTF* system, the experimental cycle of corpus research is clarified.

With the construction of corpora completed, Part 2 (Chapters 2–5) continues to pave the way for statistical processing. It centers around addressing key problems that linguists must address when engaging in the analysis of corpus data. In Chapter 2, the author shows a primacy of linguistics knowledge and the significance of devising a corpus experiment. Scientific design requires a full comprehension of the research logic which contains four crucial components: *variables*, *hypotheses*, *assumptions* and *axioms*. Specifically, a scientific research circle requires critical thinking, reframing and redefining auxiliary assumptions, variables and hypotheses. Discussions about axioms can be found in Part 3. Chapters 3–5 focus on solutions to some sociolinguistics cases. Chapter 3 pays attention to the questions of choices and baselines of appropriate linguistics models, briefly discussing the refining of a baseline for phrasal verbs by reviewing *tokens*. Chapter 4 addresses some problems in diachronic or semasiological variations, providing solutions to questions concerning how the meaning of words or expressions changes over time or with different word variations. Chapter 5 then observes variations in multi-aspects, such as genres, multi-level modelling, and choices between texts or utterances, and advocates balanced samples of the data to-be analysed.

Parts 3 and 4 are highlights of the book. Readers might, however, find it a little frustrating when getting started that, instead of hands-on software solutions, the two parts present a number of theoretical explanations along with formulae that need to be manually computed. Notwithstanding, the concern motivating this book is that researchers need to understand *exactly* what their tests are evaluating. To this end, these two parts concentrate on the introduction to basic definition and principles with scientific interpretation to concrete research cases, endeavouring to figure out the *what*, *why* and *how* of corpus statistics. Part 3, comprising eight chapters (6–13), is divided into two different sections. The first section, Chapters 6–8, concentrates on the *what* and *why*. Chapter 6 focuses on the statistical principles by means of the medium of confidence intervals, and briefly discusses the concept of possible random variation. Chapter 7 then exploits the distribution formulae in practical cases with data, plotting the intervals to understand the role of the Wilson score interval in comprehending uncertainty and confidence in the results. In Chapter 8, the author explains the relationship between confidence intervals and significance tests for three different experimental designs with Binomial variables, including 2×1 goodness of fit test, the 2×2 test for independence (the homogeneity test) and the 2×2 'independent samples from independent populations (Newcombe-Wilson)' test. The second section (Chapters 9–13) concentrates on concrete solutions to some practical test problems, ranging from comparing frequencies in the same distribution, the problem of citing and plotting a non-probabilistic property with a confidence interval, and the time series problems, to the problem of replication among different researchers' results.

To a certain extent, Part 4 is a continuation of Part 3. It emphasises some other descriptive measures, especially the size of an effect (Chapter 14) and meta-tests (Chapter 15), which may assist with experimental refinement. Chapter 8 presents solutions to two-sample z tests whose observations are expected to have the same difference and sample sizes. As complementary work, Chapter 14 pays attention to the last assessment element in significant tests, that is, effect size. This is a measurement estimating the degree to which a binomial or multinomial variable reliably predicts another, and is well-documented in the statistical literature. With both principles and relevant formulae being illustrated, Chapter 14 serves well in introducing the role of effect sizes as descriptive measures of samples in linguistics research. On the basis of Chapter 14 and existing tests, Chapter 15 offers a range of meta-tests for comparing contingency tables. It is noteworthy that in comparing tables with multiple degrees of freedom, the author shows how homogeneity gradient tests considering values of CramÉr's Φ, an effect size score with only one degree of freedom computed from a table with multiple degrees of freedom, may act as complements to the

https://doi.org/10.1017/S1351324921000413 Published online by Cambridge University Press

limited role of tests relying on the comparison of simple effect sizes. In addition to homogeneity gradient tests, another two principal classes of tests are proposed: point tests, and goodness of fit gradient tests, for further comparison of experimental results, while the common error of over-citation of critical values and *p* scores is also discussed.

After surveying common types of statistics for linguistics, Part 5 (Chapters 16–17) returns to the solutions to two methodological challenges that corpus linguists inevitably need to address: refinement of queries and overcoming sampling biases. Chapter 16 stresses the significance of data checking, and discusses the *golden rule of data*, a fundamental principle that defines the ideal population of linguistics instances, initiating the preparation of sound and complete data. It further provides complementary methods in assisting researchers to obtain the best possible estimate of the true rate of detecting and retrieving a phenomenon in language data, and in determining whether the rate of one phenomenon is greater or lesser than another. For example, to examine cases of *get* and *be* passive constructions in written and spoken parts in ICE-GB, Wallis and Mehl (in press) employed a random subsampling, choosing unequal proportions given the enormous variation in the number of occurrences of *get* (207) and *be* (9627) passive constructions. Then, through a correction to the observed proportion *p* and the employment of the Wilson score interval with subsample size *n*, they discovered that despite the rescaling of the *be* passive frequency, the two observations are significantly different. Chapter 17 mainly probes into methods for adjusting intervals for random-text samples. Two common sampling biases are proposed: convenience, and contiguous subsamples, or more accurately, *texts* in corpus linguistics. However, selected pairs from the same text would probably share more characteristics. Crucially, Chapter 17 draws on the experience of methods for adjusting the estimated variance of a sample of *cluster samples*, a similar concept to that of epidemiology. To solve this, it looks at combining techniques such as variance-weighted regression (Chapter 11) and meta-analysis (Chapter 15) into Newcombe-Wilson difference intervals and continuity corrections.

The final part (Chapters 18–19), is also composed of two chapters. Although in Chapter 7, the author has already discussed the Wilson *score interval* at length, and Chapter 11 demonstrates the logistic curve in detail, presenting how to employ the relationship between Wilson interval with logistic function for an effective logistic regression, the book has not yet examined how the Wilson *distribution* behaves and varies in different circumstances. To this end, Chapter 18 computes and plots the Wilson *distribution*, a bell-like curve which may expose how the function behaves over all possible values of α, for different values of *p* and *n*. It offers alternative distributions, and compares effects of correction on the intervals through formulae and plots. By plotting the distribution, statistical reasoning can thus be conceptualised. Finally, Chapter 19 contains some concluding remarks, briefly reviewing the main contents of the book and emphasising the author's core aim of demystifying statistical reasoning.

To summarise, one of the most prominent contributions of this book is that the author advocates abundant and scientific statistical reasoning learning, making it an excellent entry-level guide about statistics for linguistics and data processing. It builds a scientific cycle in corpus and statistical research with discussions of different experimental cases. Nonetheless, it assumes that the readers already have some basic knowledge of methods for data detecting and retrieving, statistical processing, plotting, etc., and avoids to some extent the explanation and presentation of all relevant processing steps. In this respect, the work cannot be regarded as a hands-on guide. Instead, the book focuses on teaching a fundamental and critical understanding of statistical logic, taking the view that it is the underlying statistical reasoning, rather than simple methods duplication, that matters in corpus data research. Those interested in hands-on practice could instead take a look at books such as: *Using SPSS for Windows and Macintosh (the 7th edition)* (Green and Salkind 2013); *Statistics for Linguistics with R* (Gries 2021); and the open-source *R Graphics Cookbook (the 2nd edition)* (Chang 2018).

Overall, *Statistics in Corpus Linguistics: A new approach* makes an outstanding introductory book on statistics for corpus linguistics. The lively writing style ensures high readability, especially

https://doi.org/10.1017/S1351324921000413 Published online by Cambridge University Press

for novices, while the extensive range of both theoretical and practical case studies provides a scholarly reference that will undoubtedly be valued by researchers. In a nutshell, the main target of this volume is to advocate systematic understanding of corpus research using scientific methods, and may serve as a stepping stone that leads a linguist into follow-up advanced research.

Acknowledgment. I would like to express my sincere gratitude to the review editor for the comments and language polishing.

Zheyuan Dai
Department of Linguistics, Zhejiang University,
Hangzhou 310058, P. R. China
E-mail: zydai@zju.edu.cn

References

Aarts B. (2001). Corpus linguistics, Chomsky and fuzzy tree fragments. In **Mair C.** and **Hundt M.** (eds), *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi, pp. 5–13.

Chang W. (2018). R Graphics Cookbook, 2nd Edn. Sebastopol: O'Reilly Media. Available at https://R-graphics.org/.

Green S.B. and **Salkind N.J.** (2013). *Using SPSS for Windows and Macintosh*, 7th Edn. Upper Saddle River, NH: Pearson.

Gries S.T. (2021). *Statistics for Linguistics with R*. Berlin: De Gruyter Mouton.

Wallis S. and **Nelson G.** (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery* **5**(4), 305–335. Available at https://doi.org/10.1023/A:1011453128373.

Wallis S.A. and **Mehl S.** (in press). Comparing baselines for corpus analysis: Research into the get-passive in speech and writing. In **SchÜtzler O.** and **SchlÜter J.** (eds), *Data and Methods in Corpus Linguistics: Comparative Approaches*. Cambridge: Cambridge University Press.

Yan J. and **Liu H.** (in press). Semantic roles or syntactic functions: The effects of annotation scheme on the results of dependency measures. *Studia Linguistica*. Available at https://doi.org/10.1111/stul.12177.

https://doi.org/10.1017/S1351324921000413 Published online by Cambridge University Press