Theory and Practice of Second Language Acquisition
vol. 6 (1), 2020, pp. 195–201
https://doi.org/10.31261/TAPSLA.8497

Vaclav Brezina, *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press, 2018, ISBN 978-107-12570-4, 296 pages

Curiously and notably, the study of language, in all its various forms, has always been constrained by language, which, unlike in other fields, is both an object and a tool of linguistic analysis. It should come as no surprise then, that to a researcher instilled with an acceptance of this epistemo-methodological duality, the prospect of using a nonlinguistic maths-based metric to obtain insight into how populations speak and write may appear a little daunting, if not completely disheartening. Indeed, given that statistics is understood mainly as an instrument and process of quantitative—that is, numerical—data analysis, such feelings may seem to be firmly grounded. Vaclav Brezina's book *Statistics in Corpus Linguistics: A Practical Guide* marks a departure from the traditional maths-centred presentation of statistical measures by foregrounding topics central to corpus research and language-based studies. Because the book comes with free statistical calculators, *Lancaster Stats Tools Online*, developed at Lancaster University, the focus is on understanding the principles of statistical thinking relative to linguistic datasets and variables, rather than the precise mechanics of number crunching that can be performed by the available online software. Additional materials in the form of answer keys, datasets, and teaching slides can be obtained from a companion website at Lancaster.

The volume consists of eight chapters, a final remarks section, references containing 181 entries, and a subject index. The chapters are designed as stand-alone units devoted to specific topics in research. Each chapter is structured in the same format, consisting of:

– a brief introduction explaining the aims of the chapter,

- ‘think about’ tasks which set the context for the presentation of the topic of each subsection,
- a concise note on how to report the statistics discussed in research papers,
- a practical application section which presents corpus studies offering novel insights into language use, for example: Do the British talk about the weather all the time? Or, do speech and fiction use more definite articles than general prose and academic writing? The sections illustrate queries that can be answered by applying the statistical measures commented upon in particular chapters.
- thought-provoking exercises on the procedures described in the chapter, and
- a revision section (‘Things To Remember’) and an advanced reading list for readers wishing to expand their knowledge of the chapter’s subject matter.

The adopted presentation format testifies to the author’s desire to demystify statistics and help linguists and like-minded researchers apply it in a wide range of research, including that involving smaller datasets. Frequent references to authentic studies and to their practical implications make this task more effective and intellectually captivating.

An additional strength of the book is that, alongside statistical principles, it introduces the fundamentals of corpus linguistics. In fact, it is an exhaustive compendium of expert knowledge relating to how the multiple layers of language are represented in the existing corpora and how they can be investigated and made sense of using corpus tools and statistical procedures. The information is organized in a lucid, logical and coherent manner; the style is straightforward, concise, and reader-friendly. In what follows, I provide an overview of the eight chapters by offering a critical appraisal of their contents and educational value.

Chapter 1—Introduction: Statistics Meets Corpus Linguistics (pp. 1–37) advances the perspective that “statistics in corpus linguistics is about mathematical modelling of a complex linguistic reality” (p. 5). Consequently, it introduces the scientific premises of statistics, as well as basic statistical concepts with a view to relating them to linguistic data. To this end, notions like descriptive statistics, frequency distribution, outlier, standard deviation, effect size, and many more are elaborated upon based on relevant corpus samples. The chapter also raises issues in the building of language corpora, setting up criteria for corpus representativeness and offering advice on how to avoid potential bias in the construction of corpora. As regards corpus size, the recommendation is to link it to research objectives and the investigated language point. Finally, the principles of data visualization are embarked upon. Here the author warns against the risk of data misinterpretation, which can be reduced if researchers visually familiarize themselves with the trends displayed in their results.

Throughout the chapter, as throughout the entire book, Brezina demonstrates robust theoretical knowledge of the field, combined with first-hand experience of its practical application. The latter, in particular, is revealed through various practical tips, such as that "preparing the spreadsheet in the appropriate format is as important as the statistical analysis that follows" (p. 6).

Chapter 2—Vocabulary Frequency, Dispersion and Diversity (pp. 38–65) introduces the reader to the complexities of quantitative analysis of the lexicon. The chapter opens with the statement that in corpus linguistics, a word, intuitively regarded as the prime exponent of lexico-semantic content, may be represented by four different units: tokens or running words, types, lemmas, and lexemes. Each offers different advantages and disadvantages to research and produces different error patterns. For example, performing a simple procedure such as a word count involves counting tokens, but different analysis tools operate on somewhat different notions of a token, which presents a challenge for accuracy and replicability. The next two sections discuss the measures of word frequency and dispersion, alongside average reduced frequency, a measure that combines frequency and dispersion, thus providing information on the most prominent words in a language—that is, the most frequent and evenly distributed words. The last concept addressed in the chapter is that of lexical diversity. Although specialist literature contains multiple examples of measures of lexical diversity (Malvern & Richards, 2002), most of which show sensitivity to text length, the section focusses on the select few that best illustrate the concept and/or are the most robust. Of practical value to the reader will be the remark that the mathematical equations that abound in the chapter serve an educational purpose only and that the calculations involved can be performed automatically at the companion website. What also deserves mention is the richness of contextual corpus-drawn detail that accompanies the presentation of new measures and makes their abstract mathematics more relevant to language-oriented research.

Chapter 3—Semantics and Discourse: Collocations, Keywords and Reliability of Manual Coding (pp. 66–101) continues with the subject of vocabulary, shifting focus from words in isolation to words in context. This, in Brezina's opinion, is instrumental in establishing word meanings which become apparent through the analysis of recurrent word use patterns. To understand these patterns, corpus linguistics looks at collocations and related association measures, the subject of the first thematic section of Chapter 3. Since there is no one measure to suit all research purposes, the author goes to great lengths to demonstrate the available pool of procedures, stressing the need for researchers to make informed choices from among their options. The subsequent sections expand the topic of collocations by elaborating on collocation graphs and networks as a way of visualizing word connections and their intensity, and by introducing the techniques of keywords and lockwords as metrics for conducting intercorpus comparisons. As could be expected of a book by a leading corpus linguist, the sections offer a wealth of methodological detail, including advice on choosing adequate corpora, dealing with absent words and applying the right statistical tests. Concerning the latter, the interested reader will find here a criticism of the traditional log-likelihood statistic and a recommendation to use the more robust simple maths parameter (Kilgariff, 2009). The final theme of the chapter is inter-rater agreement, an issue in tests that require subjective judgement and evaluation, such as deciding on a word's positive or negative connotations. In conclusion, considering the breadth and depth of the information it provides, coupled with the clarity of presentation, the chapter is a comprehensive resource for novice and experienced researchers alike.

Chapter 4—Lexico-grammar: From Simple Counts to Complex Models (pp. 102–138) narrows the focus down to lexico-grammatical features. In corpus linguistics, the 'label' refers to specific constructions or expressions, such as articles or passives. The chapter compares and contrasts the two research designs used in analyses of lexico-grammar—the whole corpus design and the linguistic feature design—elaborating on the explanatory value of their output. It then goes on to illustrate the application of simple cross-tabulation and chi-squared tests, and outlines the conditions for their use and potential weaknesses, such as sensitivity to sample size. A useful tip for researchers is that with corpus datasets, which are usually massive, the expected frequency assumption tends to be easily met. For more complex computations involving multiple heterogenous—that is, categorical and scale—variables, the recommendation is to run logistic regression and build a model configuring the variables concerned. The procedure and stages inherent in the process are meticulously described in the chapter. Nevertheless, Chapter 4, gives the impression of being overly abstract and mathematical, which may be a challenge to the unaccustomed reader. On the other hand, since regression models are popular measures with enormous explanatory power, the mathematics may be necessary to help researchers understand the perspective on language data that logistic regression provides. Indeed, the author himself offers a reminder that it is essential to understand the basic principles of the test and the interpretation of the output. The computation can be performed automatically by computers.

Chapter 5—Register Variation: Correlation, Clusters and Factors (pp. 139–182) examines the topic of the relationships that hold between linguistic variables in different registers and genres. The most straightforward relationship is that of correlation, which is represented by Pearson's and Spearman's correlations. The chapter explains both with clarity and in detail. The author warns against placing too much trust in statistical significance because in the case of correlation it is directly related to the number of observations (p. 144). Therefore, the correlation coefficient should be reported together with the confidence interval. Linguistic variables may also function as defining features or descriptors. Statistics offers a technique called hierarchical agglomerative cluster analysis, which visualises patterns of category (group) membership based on two features (descriptors), and as such is demonstrated in the next subsection of the chapter. When the number of the variables to consider rises, as may be the case with register comparisons, it is necessary to use a variant of cluster analysis in the form of multidimensional analysis. It is by far the most complex of the techniques discussed in the book, chiefly on account of the necessity to group (load) tens or even hundreds of descriptors (variables) into factors representing more general features. Here, once again, Brezina shows his masterful grasp of statistics and research methodology as the procedure is described step-by-step with multiple examples and authentic datasets (Biber 1988), as required by each stage.

Chapter 6—Sociolinguistics and Stylistics: Individual and Social Variation (pp. 183–218) uses the notion of style and stylistic variation in speech and writing to set the context for inter- and intragroup (speaker) comparisons. The chapter begins with an evaluation of Labov's and Biber's approaches to individual and social variation and their implications for the identification of variables in research. It then embarks on an analysis of whether the speaker's gender is related to the frequency of use of personal pronouns. The statistical techniques recommended for the process include Welch's independent samples t-test, which compensates for unequal variances (one of the t-test's assumptions). As has been the custom in the present book, the procedure is explained with replicable clarity and precision. As an additional bonus, often overlooked by older statistics textbooks, the chapter comments on the need to calculate an effect size and offers an interpretation of the measure. In a similar vein, other related statistical tests are discussed, including one-way ANOVA, post-hoc tests and the nonparametric Mann-Whitney U test and the Kruskal-Wallis test. It cannot escape notice that these tests are the classics of inferential statistics and as such are household terms not only in corpus research but also in other fields such as second language acquisition. The next issue taken up by the author is correspondence analysis, whose output visualizes the linguistic characteristics of individual speakers in a manner similar to cluster analysis discussed in Chapter 5. In turn, a reader with an interest in forensic linguistics will find that mixed-effects models have the capacity to identify the author of a particular text based on the individual's choice of words. A special merit of the chapter is that by sifting through data and performing analyses that only a few years ago seemed nearly impossible, it reveals the enormous exploratory potential that the application of statistical methods to linguistics may create.

Chapter 7—Change over Time: Working Diachronic Data (pp. 219–256) looks at ways of analysing linguistic change in historical or diachronic corpora. Since such data tend to be limited to written records and are often biased towards certain genres or types of author, researchers are advised to consider their options carefully and make “the best use of bad data” (p. 222). The chapter presents a handful of techniques with a focus on probing stability and change over time as significant variables in the evolution of language. They include analyses of the percentage change of variables and the nonparametric bootstrapping test, which allows comparisons of two corpora representing different points in time through multiple resampling of the available material. The other recommended procedures involve different forms of visualisation such as cluster analysis, the peaks and troughs technique and its extension called usage fluctuation analysis. The chapter ends with a resumé of the author’s explorations into the realm of seventeenth century colour terminology which demonstrates implementation of the ideas discussed in the chapter.

Chapter 8—Bringing Everything Together: Ten Principles of Statistical Thinking, Meta-analysis and Effect Sizes (pp. 257–282) is a summary of the rules and guidelines regarding good practices in statistical analysis. It opens with a list of ten principles that ensure the precision and rigour of research findings. These include attention to detail during the data processing stage, informed choices of statistical procedures coupled with transparency of their presentation, reporting effect sizes in addition to *p*-values, and visualizing data to highlight patterns, to mention just a few. The author also stresses the importance of pooling findings together to obtain a global perspective on a specific research query. This can be done through meta-analysis, which synthesizes research results by comparing the effect sizes of compatible studies and calculating a summary effect. The discussion ends with advice on how to use, interpret and report the various effect sizes introduced in the book. Perhaps one of the author’s most fitting comments on good practice in research is that found in ‘Final Remarks’ (pp. 283–284). It reads as follows: ‘Students often ask me what the best statistical test is to use with corpora […] I usually respond: in many cases, the most powerful statistical technique is common sense’ (p. 284).

Without doubt, Vaclav Brezina’s volume, together with the companion website, is a powerful resource for linguistic research. The point I have been trying to make in this review is that, as a resource, it is also flexible and versatile because the principles of analysis it lays out so competently can be applied to any collection of texts, including those by second/foreign language learners and multilinguals. Further, since the majority of bibliographical sources referred to in the book were published after the year 2000, the book provides a most recent state-of-the-art review on the subject. A potential lacuna is a lack of information on how to process data prior to statistical analysis. This is essential in the light of the fact that many statistical tests require specific variables that need to be extracted from annotated corpora, using dedicated software. Also, some of the calculators at the companion website require training and are not intuitively easy to work with. Overall, however, the volume is an unrivalled theoretical and practical toolbox for researchers wishing to understand research reports, and construct and analyse their own datasets, and as such should be a top entry on each applied linguist's reading list.

https://orcid.org/0000-0002-9478-7689

Jolanta Latkowska
University of Silesia in Katowice, Poland