Journal of Quantitative Linguistics
2010, Volume 17, Number 1, pp. 74–79
DOI: 10.1080/09296170903395981

BOOK REVIEW

Johnson, Keith, *Quantitative Methods in Linguistics*. Malden, MA: Blackwell, 2008. ISBN: 978-1-4051-4424-7, 277 pp.

Reviewed by Emmerich Kelih and Ján Mačutek
Institute of Slavistics, University of Graz

The book under review comprises seven chapters and is designed as an introduction to quantitative methods in linguistics. The target audience – although this is not explicitly mentioned – are students of linguistics and others who are interested in the application of quantitative/statistical methods in linguistics in a "... relatively representative and usable introduction to quantitative research across many different subdisciplines within linguistics" (p. x).

After a short discussion of the author's motivation for writing such a textbook ("... not happy with the available [but not further specified: EK, JM] textbooks" (p. 1)) the conceptual design of the book is presented. This consists of three parts:

(1) Presentation of basic statistical methods.
(2) Some discussions on the practical use of statistical methods in different subdisciplines (phonetics, sociolinguistics, psycholinguistics, historical linguistic and syntax).
(3) Instructions for the open-source statistical software package *R*.

All seven chapters of the book are accompanied by exercises, which give the reader an opportunity to practise the statistical methods discussed.

In this review we provide a short summary of each chapter, and some critical remarks regarding the theoretical and methodological approach provided in the book. It should be noted that all the linguistic problems and data treated in the textbook are either from published sources or from the author's own linguistic experiments.

Chapter 1, "Fundamentals of quantitative analysis", starts with some basic terms of descriptive statistics (frequency distributions, measures of central tendency, measures of dispersion), including different kinds of scales of measurement. As early as the first chapter, different types of distributions (normal bell-shaped, bimodal, U-shaped distributions, etc.) are discussed. Special attention is paid to the normal distribution in statistics generally, but without the requisite linguistically motivated critical distance.

The second chapter, "Patterns and tests", begins with an extended discussion of the well-known statistical concept of random samples and general population, with reference to the central limit theorem; that is some standard textbook information about analytical statistics is provided. Different types of hypothesis testing and methods of exploring relationships among variables (i.e. correlation and regression analysis) are presented.

The third chapter, "Phonetics", the first chapter with a special focus on linguistic problems, illustrates in a relatively simple way different tests of significance (*t*-test for independent samples and paired comparisons) and multivariate methods (multiple regression, principal components analysis). These methods are mainly based on Voice-Onset time measurements, performed by the author.

The fourth chapter, "Psycholinguistics", deals with a small number of psycholinguistic phenomena; namely, some experimental data from word-perception studies and influence factors are used to illustrate methods of the so-called analysis of variance (ANOVA).

The fifth chapter, "Sociolinguistics", is mainly dedicated to sociolinguistic problems, e.g. data about the different phonetic realization of phonemes by speakers, dependent on age and sex, etc., are discussed. These data are used for the discussion of statistical methods based on contingency tables and logistic regression. Interestingly, when discussing the binomial distribution in this chapter, the author uses data of an exceptionally non-linguistic nature (opinion poll results).

Different clustering methods and the method of multidimensional scaling – that is, once again methods of multivariate statistics – are

presented in the sixth chapter on "Historical linguistics", where mainly examples from studies dealing with the (phonological) similarity and historical relatedness of languages are discussed.

The final seventh chapter, "Syntax", is from a statistical point of view dedicated to methods already mentioned in the previous chapters, e.g. this chapter provides the reader with a deeper understanding of linear/non-linear regression analysis and ANOVA, based on data from experimental/empirical studies of speaker's acceptability judgments. The text book ends with an appendix, a detailed subject index and references.

In summary, we are dealing with a well-structured and balanced textbook of quantitative methods in linguistics, which also provides in a satisfactory way recent and attractive possibilities of quantitative linguistic analyses as an easily understandable explanation of adequate statistical methods. The textbook is perhaps less successful in certain other regards. The limited relevance of classical statistical methods (based on the normal distribution) in linguistics is not treated satisfactorily. The author presents only well-known statistical "textbook knowledge" on "representative" samples, general population, etc. But in linguistics there is no evidence for the validity of a simple "representativeness-idea". Furthermore, it must be emphasized that this problem is directly connected with the validity of the normal distribution: Except for some psycho- and sociolinguistic data perhaps, linguistic data *are not normally distributed*. These problems, termed by Köhler and Altmann (2005) and Köhler (2005) "inferential problems" (representativeness, homogeneity, normal distribution of the random variable, homoscedasticity), leads to substantial limitations of the applicability of many statistical tests and statistical methods based on the normal distribution. In fact, for linguistic data extreme skewness is typical and, consequently, much more attention must be paid to the choice of adequate statistical methods (e.g. non-parametric) than in other sciences. Without a discussion of these central problems for the application of quantitative methods in linguistics, a textbook remains incomplete. Moreover, some further weak points of the book must be mentioned.

Page 3, Section 1.1: The author does not write a single word on the problem of (often not easy) ways from a formulation of a qualitative linguistic hypothesis to its quantification and back to its qualitative interpretation. Following Altmann (1973) and Köhler and Altmann (2005), investigations in the field of quantitative linguistics should follow a scheme of five steps: firstly, linguists must set up an empirically relevant

and testable hypothesis; secondly, the hypothesis has to be "translated into the language of statistics"; and thirdly, appropriate statistical methods have to be found. Fourthly, based on the statistical results and significance tests, the hypothesis must be refused or accepted. Finally the hypothesis, i.e. the examined linguistic question, must be interpreted. Following these steps, it becomes clear that statistics and quantitative methods assume the role of an auxiliary means of deductive linguistic research.

Page 12: Most often, the source of variability is not only some type of random error in observations/measurements: language itself is non-deterministic. Page 13: The area under every density function is 1; it is not a property specific to the normal distribution. Pages 15–24: Section 1.5: Normality, as already mentioned, is not very normal in linguistics. One should consider it rather an exception, or a very rough approximation (which can hide very interesting aspects of investigated phenomena). Next, the Q-Q plot is a good first hint but one cannot rely on "optical testing". In *R* exact normality tests are also implemented, e.g. the Shapiro-Wilk test (see shapiro.test in *R*). Page 26: The formula for the mean is wrong. On page 25 we read that to obtain the mean one should "sum the data values and then divide by the number of values in the data set" – completely correct, but if the values are denoted $x_0, x_1, \ldots, x_n$, then there are $n + 1$ values, hence the sum must be divided by $n + 1$, not by n. Page 38: The central limit theorem is doubtlessly a very useful tool and it is often applied. However, its most simple proof makes use of the apparatus of characteristic functions and one is expected to have basic knowledge of complex analysis to be able to understand it. What the author describes (he tries to provide a kind of a constructive proof) is just an indication of how things develop. But the central limit theorem, true to its name, is a limit theorem – it deals with infinity. If one wishes to prove the theorem in the way suggested by the author, one must be patient enough to perform an infinite number of steps. Alternatively, a more realistic possibility is to call the author's approach a hint, an illustration or something of that kind, definitely not "the beginning of the proof". Moreover, the central limit theorem is true for most known distributions. Here only one of them is arbitrarily chosen, which cannot provide a general proof. Page 42: The relation

$$\frac{d}{dx}F(x) = f(x)$$

is true only at the points at which the distribution function is differentiable. Page 112: "ANOVA ... assumes that error values are normally distributed", but "ANOVA is actually pretty robust in the face of assumption violations". Indeed it is. Nevertheless, if the author wants to recommend using ANOVA without thinking of (and testing) its assumptions, this can be a rather dangerous recommendation.

It goes without saying that not all linguistic levels at which quantitative methods are applicable and already used can be presented in a textbook with limited space. But as the author already points out, in the future, examples and case studies from other linguistic disciplines should be included. Perhaps not only should a chapter on language typology be added, as the author writes (p. ix), but also some main issues from so-called synergetic linguistics and general quantitative linguistics (cf. Köhler, 2005). In these linguistic disciplines quantitative methods are also understood as an appropriate tool for empirical analysis, but unlike the epistemological positioning in the textbook under review, synergetic/quantitative linguistics is defined as a deductive science, which mainly investigates functional, distributional and developmental laws by means of statistics (cf. Altmann, 1993, 1997) and Köhler and Altmann (2005). So in fact, the integration of some examples from synergetic/quantitative linguistics would enrich the textbook. But as the author correctly points out – unfortunately scattered throughout the textbook and only in a rudimental way – the era of a deterministic viewpoint in linguistics is finally over (p. 247). Linguistics is clearly on the way to becoming a fully-fledged member of the scientific community, in which it is quite normal to use quantitative methods to explore *a priori* defined hypotheses and to provide adequate interpretations of the results.

Overall, though it has its shortcomings, the book provides a useful introduction to quantitative methods in some fields of linguistics. Moreover, a thorough explanation of many basic (and some not so basic) features of *R* is given. For a linguist, the ability to work with this software and to solve practical problems can be of great help in gaining statistical insight. However, the book could be much better. One of its (almost certainly unintended) achievements is that it silently highlights the need for co-operation. Co-operation with (not only American) linguists reduces the risk of rediscovering already known facts (the references lack many important books and papers by European and Japanese linguists) and of repeating "traditional" methodological mistakes. Without co-operation with mathematicians, a linguist often achieves inexactness with

exact methods (and a mathematician writing such a book without co-operation of linguists would most probably produce absolutely exact nonsense). There are very few linguists, and probably no mathematicians, who can write alone a textbook on quantitative methods in linguistics without mistakes that can easily be avoided if the team of authors is "mixed".

All in all, the book under review can be recommended to students and to all linguists who are interested in obtaining a quick overview and easily understandable presentation of some important and useful statistical/quantitative methods in linguistics. In addition, of course, V. Altmann and G. Altmann (2008), Baayen (2008), and Tuldava (1995, 1998) should be used for a more precise and representative overview over the problems and the effectiveness of quantitative methods in linguistics.

REFERENCES

Altmann, G. (1973). Mathematische Linguistik. In W. A. Koch (Ed.), *Perspektiven der Linguistik* (pp. 208–232). Stuttgart: Kröner.

Altmann, G. (1993). Science and linguistics. In R. Köhler & B. B. Rieger (Eds), *Contributions to Quantitative Linguistics* (pp. 3–10). Dordrecht: Kluwer.

Altmann, G. (1997). The art of quantitative linguistics. *Journal of Quantitative Linguistics*, *4*, 13–22.

Altmann, V., & Altmann, G. (2008). *Anleitungen zu quantitativen Textanalysen. Methoden und Anwendungen* [*Studies in Quantitative Linguistics*]. Lüdenscheid: Ram-Verlag.

Baayen, R. H. (2008). *Analyzing Linguistic Data*. Cambridge: Cambridge University Press.

Köhler, R., & Altmann, G. (2005). Aims and Methods of Quantitative Linguistics. In G. Altmann, V. Levickij & V. Perebejnis (Eds), *Problemy kvantytatynoï lingvistyky* [*Problems of Quantitative Linguistics*] (pp. 12–41). Černivci: Ruta.

Köhler, R. (2005). Synergetic linguistics. In R. Köhler, G. Altmann, R. G. Piotrowski (Eds.), *Quantitative Linguistik/Quantitative Linguistics* [*Handbücher zur Sprach- und Kommunikationswissenschaft, 27*] (pp. 760–774). Berlin: De Gruyter.

Tuldava, J. (1995). *Methods in Quantitative Linguistics* [*Quantitative Linguistics, 54*]. Trier: Wissenschaftlicher Verlag Trier.

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie* [*Quantitative Linguistics, 59*]. Trier: Wissenschaftlicher Verlag.