




<https://doi.org/10.22363/2687-0088-30209>

Book review

**Review of Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York/Oxon, Routledge.  
ISBN 9781138589384 ISBN 9780429491696 (eBook)**

**Irina PRIVALOVA<sup>1</sup>   and Mariia KAZACHKOVA<sup>2</sup> **

<sup>1</sup>*Kazan (Volga region) Federal University, Kazan, Russia*

<sup>2</sup>*Moscow State Institute of International Relations (University), Moscow, Russia*  
 [angladkova@gmail.com](mailto:angladkova@gmail.com)

**For citation:**


Privalova, Irina & Mariia Kazachkova. 2022. Review of Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York/Oxon, Routledge. *Russian Journal of Linguistics* 26 (2). 550–557. <https://doi.org/10.22363/2687-0088-30209>

Рецензия

**Рецензия на книгу  
Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York/Oxon, Routledge.  
ISBN 9781138589384 ISBN 9780429491696 (eBook)**

**И.В. ПРИВАЛОВА<sup>1</sup>  , М.Б. КАЗАЧКОВА<sup>2</sup> **

<sup>1</sup>*Казанский федеральный университет, Казань, Россия*

<sup>2</sup>*Московский государственный институт международных отношений (университет), Москва, Россия*  
 [angladkova@gmail.com](mailto:angladkova@gmail.com)

**Для цитирования:**

Privalova I., Kazachkova M. Review of Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York/Oxon, Routledge. *Russian Journal of Linguistics*. 2022. Vol. 26. № 2. P. 550–557. <https://doi.org/10.22363/2687-0088-30209>

In present-day scholarship, there exist various approaches to the study of the material included in language corpora. The majority of the corpora studies are focused on practical issues of foreign language acquisition and provide samples of classroom activities that engage corpus material (Friginal 2018, Perez-Paredes 2020, Timmis 2015). There are also works devoted to particular linguistic issues,



for instance, the use of corpus linguistics in grammar analysis (Jones & Waller 2015), stylistics, writing and thought presentation (Semino & Short 2011) or special aspects of vocabulary (Szudarski 2017). The study of corpus linguistic data within the frame of non-linguistic disciplines is a rare phenomenon. One of the most successful attempts is an in-depth analysis of corpora use in sociolinguistics. This has become possible due to easier access to data in different languages through websites, social networking sites and blogs (Friginal & Hardy 2014).

The statistical approach to the study of linguistic corpora is of utmost importance since statistical methods are essentially helpful in assessing the representativeness of the language material and the reliability of the obtained results. In addition, statistical methods have proved to be quite effective in building statistical models and making forecasts. “Corpus linguistics is a powerful quantitative methodology, which heavily relies on frequency data and statistical procedures. It is difficult to talk about corpus linguistics without mentioning statistic measures based on frequency and distribution” (Cunxin 2020: 379). No wonder that we have recently witnessed the publication of a number of monographs debating the possibilities of statistical analysis. A bird’s-eye view on basic statistics for corpus linguistics including both inscriptive and inferential techniques was presented almost two decades ago by Oakes in his book “Statistics for Corpus Linguistics” (Oakes 1998). It is also worth mentioning the edition “Quantitative Corpus Linguistics with R. A Practical Introduction” (Gries 2016), which demonstrates how to process corpus-linguistic data with the open-source programming language and environment R. Along with data and text processing, the author dwells upon basic aspects of statistical analysis and visualization. Vaclav Brezina’s monograph “Statistics in Corpus Linguistics: A Practical Guide” (Brezina 2018) was the first attempt not only to provide the theoretical underpinning of statistical analysis but also to consider some practical examples of statistical techniques application. This volume contains some self-study material on a special companion website with exercises, keys and datasets. Cunxin Han, who carried out the analysis of all the three above-mentioned monographs on statistics in linguistic corpora, writes: “Brezina’s book is a timely contribution because hands-on books are rarely found in the literature. The first book dedicated to this field “Statistics for Corpus linguistics” is already 20 years old while a more recent book “Corpus linguistics and statistics with R”, caters more to the needs of advanced researchers” (Cunxin 2020: 379).

Amidst the body of similar studies, the monograph under review “Statistics in Corpus Linguistics: A New Approach” (2021) by Sean Wallis stands out. This volume is in a league of its own since it unites the achievements of statistics and corpus linguistics. The title of the book declares the *new approach*, and it is indeed presented in this edition. In the preface, the author poses the question: “Why do we need another book on statistics?” which he answers in the following way: “This book arose from the realisation that conventional approaches to the teaching and discussion of statistics – in the field of linguistics at least – are not working” (VIII).

It is worth saying a few words about the author of this book and about the history of its creation. Professor Sean Wallis is the principal research fellow at the Department of English Language and Literature at the University College London and the Head of the first Corpus Linguistics research centre established in Europe. Professor Sean Wallis has been preoccupied with compiling corpora for more than 25 years and he is one of the leading experts in building corpora whose work has international acclaim. Interestingly enough, this monograph (as the author reveals) was written in stages over ten years. Some early versions of two chapters appeared more than ten years ago and were afterwards refined; later new chapters were added. As such, this book is a multi-year and well thought-out research.

The narration of the book “Statistics in Corpus Linguistics: A New Approach” is rather well-elaborated as the book consists of six parts, nineteen chapters, preface, glossary, reference list, and index. Each chapter in its turn has sub-chapters and paragraphs. The following practical issues are put in the spotlight: the constraint of a research question, the correct employment of confidence intervals, the selection of optimal significance tests, the effect of variables on each other, the estimation of distribution patterns similarity, and comparison of the results of two experiments (2). Some new issues are debated in this book, specifically, the role of ideal binomial and normal distributions in various tests as well as the possibilities of the Wilson interval compared to the Gaussian (normal) interval. The requirements to the linguistic experiment that produce valid language data are also described. In the section about the needs of linguists in statistics Wallis points out that the main problem is the correct choice of the test – whether it should be a chi-square or log-likelihood. Also, it is stated that the relevance of the obtained data matters as well as the effect of one variable on another.

A remarkable feature of the book is its discursive style; the author seems to be in a dialogue with his readers. For example, in the Preface the titles of paragraphs are formulated as questions: “Why do we need another book on statistics?”, “Why is statistics difficult?”, “What do linguists need to know about statistics?” (XIII–XXII). However, the introduction that looks like a conversation should not create an illusion of an easy read, because the author poses some research questions that cannot be tackled without any basic knowledge in statistics: how to employ confidence intervals to correctly measure the size of the effect of one variable on another, how to estimate the similarity of distribution patterns; how to evaluate whether the results of two experiments significantly differ, and so on. A big number of new special terms might scare off amateurs in statistics. To prevent this from happening, the author provides a Glossary (pp. 329–341) wherein the explanation of the most important concepts can be found. There are definitions of some universal concepts such as “algorithm – a complete description of a computational process” (p. 329) or “an axiom – a rule or principle of a mathematical system” (p. 330) or “research question – a general question motivating a piece of research that may require refinement in order to be translated into testable hypothesis” (p. 366). In addition, there are many terms related to statistics, for instance:

“Student’s  $t$  test – a comparable test to the  $z$  test, properly for Real data (usually termed on a Ratio scale) but sometimes used for Interval data” (p. 340).

Part 1 “Motivations” discusses general considerations on motivations, along with the issues of experimental design and data collection in corpus linguistics. Wallis clarifies the principles of material collection for corpora building. At the very beginning, the author emphasizes that the number of words does not necessarily lead to information quality. The size of the corpus should be in agreement with its annotation richness. In this section, the readers should pay attention to three important issues. First, a clear demarcation between such notions as “corpus” and “dataset”: “In the most general sense, corpora are simply collections of language data processed to make them accessible for research purposes. In contrast to experimental datasets sampled to answer a specific research question, corpora are sampled in a manner that – as far as possible – permits many different types of research question to be posed. Datasets extracted from corpora are not obtained under controlled conditions but under ‘naturalistic’ or ‘ecological’ ones” (p. 4).

Second, the author distinguishes corpora that include written texts and corpora of spoken data. A major part of the corpora is drawn from written sources, “...a corpus of spoken data, ideally in the form of recordings aligned with orthographic transcriptions” (p. 5). When data come from real world sources rather than controlled laboratory conditions, they are more informative and relevant. Speech corpora represent a pioneering approach to make the recordings of discourses, dialogues and responses. It is an approach to register the multimodal essence of the language with sound evidences and prosodies (Svenja & Carter 2013). Linguistic data obtained from psycholinguistic experiments can be included in language corpora; however, these experiments must be carried out with little (or no) guidance from the experimenter. There are essentially three distinct classes of empirical evidence that may be obtained from any linguistic data source, and they are: factual evidence, frequency evidence and interaction evidence (p. 6). Wallis argues that the same statistical methods can be applied to different types of corpora in linguistics, including text corpora, speech corpora, lexical corpora and inter-lingual corpora. He considers monolingual, bilingual and even trilingual corpora, which are widely used as electronic dictionaries, as examples for statistical analysis. The necessity to use statistical methods is also determined by the fact that modern linguistic corpora are a repository of big data. The successful development of corpus linguistics goes hand in hand with the development of new computer technologies. Digitalisation of the language provides for the centralisation of information in many forms, one of which is language corpora. Therefore, linguistic corpora can be viewed as examples of a structured representation of big data, the validity of which can be proved by the application of statistical techniques.

Part 2 “Designing Experiment with Corpora” presents an analysis of the statistical methods that may be applied in corpus experiments. Since many of the studies in corpus linguistics come from literary disciplines, the researchers may be

unfamiliar with mathematics and statistics methodologies. The author asserts that experimental methods in corpus linguistics have their specifics and require a creative approach. Corpus linguists should be careful about controlled ‘laboratory’ experiments that use stimuli, ‘cues’, or artificial conditions to encourage particular behaviours: “The dominant trend in corpus linguistics is to build ever-larger ‘flat’ tagged corpora and employ greater reliance on computation” (29). This part analyses such techniques as obtaining data, extracting data, visualising proportions and applying testing (the Chi-Square test). An example of a linguistic interaction experiment is considered on pp. 40–42. In the chapter “The Vexed Problem of Choice”, the author underlines the axiomatic character of models in sociolinguistics and cognitive linguistics research and the non-axiomatic one in corpus linguistics. It would be an exaggeration to say that experimental models in corpus linguistics are limited to mere ‘counting surface phenomena.’ Linguistic choice corpus research requires the inference of the counterfactual. Alongside what participants wrote or said, the researcher needs to infer what they could have written or spoken instead. To take a simple example, consider the study of *that*-omission in relative clauses, as in *The man [that] I saw*. We must be able to reliably identify ‘zero-relative’ clauses (p. 51). As for exposure rates, the author affirms that corpora are exceptional resources for estimating overall likelihood of readers or hearers encountering a form. There are different formulas that present exposure rate ( $p(x \text{ I word}) = f(x) / f(\text{words})$ ) and choice rate ( $p(x \text{ I } X = / f(X)$ ) (51).

Before the readers pass on to Parts 3, 4 and 5, we would strongly recommend them to take a closer look at the definitions in the Glossary. Special statistical terms are in abundance in Parts 3, 4 and 5, and it is challenging to read the text without the knowledge of such terms as: ‘distribution’, ‘interval’, ‘probability’, ‘variable’, ‘baseline’, ‘contingency’, ‘uncertainty’, etc. In Part 3 “Confidence Intervals and Significance Tests”, Wallis makes an introduction into inferential statistics and contemplates on the ‘naïve knowledge’ that people get through experience. The same is true about the knowledge in statistics which the author calls ‘naïve statistics’. He gives the definition of a probability and distinguishes three types: the observed probability (or proportion), the ‘true’ population probability and the probability that an observation is unreliable (p. 98). Also, the phenomena of binominal, normal and skewed distributions are presented in the possible applications towards texts. Optimum methods of calculation of linguistic data are presented on p. 212.

In Part 4 “Effect Sizes and Meta-Tests”, Wallis compares different versions of the same experiment in order to see how a change to the experimental design may influence the results. He also examines the effect of an experimental design upgrading on reported results. The author introduces such notions as ‘point test’ and ‘multi-point test’ for contrasting the distribution of linguistic data across a dependent variable in homogeneity tables, as well as ‘gradient test’ methods for comparing sizes of effect in homogeneity tables, commencing with intervals and tests with a single degree of freedom. He also describes the application of the test

for Cramer's effect sizes larger table. One more novelty is the consideration that linguistic variables may often be measured as Binomial proportions expressing the probability that in a random case drawn from a sample the people might find a particular linguistic phenomenon (p. 231).

Part 5, "Statistical Solutions for Corpus Samples", addresses particular problems, such as conducting research with imperfect data and adjusting intervals for random-text samples. This chapter considers situations and statistical principles that can be used in their relation to linguistic matter. The author considers variations of observed proportions between text subsamples utilizing two different models: one that analyses each text as a random sample, and another that examines the distribution of actual subsample scores. The discussion of the possibilities of the new method brings the author the conclusion that the 'Binomial' per-text distribution is really the sum of multiple Binomial distributions, one for each sample size (pp. 278–279).

Part 6 "Concluding Remarks" shows how to calculate distribution curves, for instance, how to plot the Wilson distribution and Clopper-Pearson distribution. The Wilson score interval is a member of a class of confidence intervals characterising expected variation about an observed Binomial proportion. The author concludes that the entire point of a statistical method is to understand the implications of their data. "Research concerns the structure of the physical world: in linguistics, the structure of language" (p. 314). The chapter ends with a brief summary of the author's ideas.

In conclusion, it is necessary to note that the availability of corpora and the technological advancements of corpus tools have recently increased dramatically. At present, digitalisation of the language, globalisation of information processes and complication of text models require new methods for studying linguistic material. In order to show groundbreaking results, scientists have to go beyond the borders of one linguistic discipline, and the monograph "Statistics in Corpus Linguistics: A New Approach" by Sean Wallis shows how this can be done. It demonstrates a new non-trivial approach towards statistical methods and the way to apply it when studying quantitative and qualitative linguistic variables, such as a specific lexical unit in various syntactic constructions, the distribution of word length, the distribution of sentence length, and the distribution of specific vocabulary in the text. No doubt, this book will not go unnoticed by "pure" linguists since the author claims it is "...accessibly written to those with little to no statistical background" (i). Indeed, this book is written in such a way that a set of specific knowledge and terms would not scare away a beginner. Moreover, researchers of a "new generation" working in the field of corpus linguistics cannot do without any basic information in the field of programming and statistics. Thus, the advantages of the Chi-Square testing or t-Student's testing in their application towards corpus linguistics are obvious. By and large, the monograph "Statistics in Corpus Linguistics: A New Approach" by Sean Wallis is not a book for one-time reading, rather it is a guide that scientists will refer to on a regular basis when exploring some kind of linguistic material presented in corpora.

## Acknowledgements

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

## REFERENCES

- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781316410899>
- Cunxin, Han. 2020. Statistics in Corpus Linguistics: A Practical Guide. *Journal of Quantitative Linguistics* 27(4). 379–383. <https://doi.org/10.1080/09296174.2019.1646069>
- Friginal, Eric & Jack Hardy. 2014. *Corpus-Based Sociolinguistics. A Guide for Students*. N.Y.: Routledge. <https://doi.org/10.21283/2376905X.2.32>
- Friginal, Eric. 2018. *Corpus Linguistics for English. Teachers. Tools, Online Resources, and Classroom Activities*. N.Y.: Routledge. <https://doi.org/10.4324/9781315649054>
- Gries, Stefan Th. 2016. *Quantitative Corpus Linguistics with R. A Practical Introduction*. N.Y.: Routledge. <https://doi.org/10.4324/9781315746210>
- Jones, Christian & Daniel Waller. 2015. *Corpus Linguistics for Grammar. A Guide for Research*. N.Y.: Routledge. <https://doi.org/10.4324/9781315713779>
- Oakes, Michael. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University press. ISBN 0-7486-103204
- Perez-Paredes, Pascual. 2020. *Corpus Linguistics for Education. A Guide for Research*. N.Y.: Routledge. DOI: 10.4324/9780429243615
- Semino, Elena & Mick Short. 2011. *Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing*. N.Y.: Routledge. <https://doi.org/10.4324/9780203494073>
- Szudarski, Pawel. 2017. *Corpus Linguistics for Vocabulary. A Guide for Research*. N.Y.: Routledge. <https://doi.org/10.4324/9781315107769>
- Svenja, Adolphs & Ronald Carter. 2013. *Spoken Corpus Linguistics. From Mono-modal to Multimodal*. N.Y.: Routledge. <https://doi.org/10.4324/9780203526149>
- Timmis, Ivor. 2015. *Corpus Linguistics for ELT. Research and Practice*. N.Y.: Routledge. <https://doi.org/10.4324/9781315715537>
- Wallis, Sean. 2021. *Statistics in Corpus Linguistics: A New Approach*. N.Y./Oxon: Routledge. <https://doi.org/10.4324/9780429491696>

## Book review history:

Received: 20 December 2021

Accepted: 25 February 2022

## Bionotes:

**Irina V. PRIVALOVA** is Doctor Habil. of Philology, a Leading Research Fellow of the Research Laboratory “Text Analytics” at the Institute of Philology and Intercultural Communication of Kazan (Volga Region) Federal University. Her research interests embrace psycholinguistics, intercultural and mass communication, as well as corpus linguistics.

## Contact information:

Kazan (Volga Region) Federal University  
building 33, 2 Tatarstan street, Kazan, 420021, Russia  
e-mail: ivprivalova@mail.ru  
ORCID: 0000-0002-7740-2185

**Maria B. KAZACHKOVA** is an Associate Professor at Moscow State Institute of International Relations (University). Her research interests include corpus linguistics, text analytics, and discourse studies.

**Contact information:**

Moscow State Institute of International Relations (University)

3 Novo-Sportivnaya st., Odintsovo, 143071, Russia

*e-mail:* mbkazachkova@yandex.ru

ORCID: 0000-0002-0357-3010

**Сведения об авторах:**

**Ирина Владимировна ПРИВАЛОВА** – доктор филологических наук, ведущий научный сотрудник НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета. Ее исследовательские интересы включают психолингвистику, межкультурную и массовую коммуникацию, корпусную лингвистику.

**Контактная информация:**

Казанский федеральный университет

420021, Казань, ул. Татарстан 2, здание № 33, комната № 46

*e-mail:* ivprivalova@mail.ru

ORCID: 0000-0002-7740-2185

**Мария Борисовна КАЗАЧКОВА** – кандидат филологических наук, доцент Московского государственного института международных отношений (университета). Ее научные интересы включают корпусную лингвистику, аналитику текста и исследование дискурса.

**Контактная информация:**

Московский государственный институт международных отношений (университет)

Россия, 143071, Одинцово, ул. Ново-Спортивная, д. 3

*e-mail:* mbkazachkova@yandex.ru

ORCID: 0000-0002-0357-3010