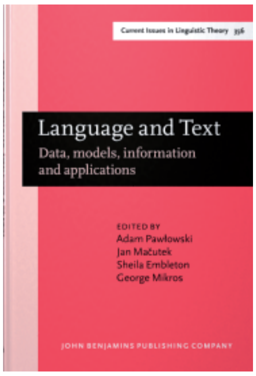
RECENZJE

ROCZNIKI HUMANISTYCZNE
Tom LXX, zeszyt 8 – 2022

Adam Pawłowski, Ján Mačutek, Sheila Embleton, & George Mikros, editors. *Language and Text. Data, models, information and applications*. John Benjamins, 2021, 280 pp. ISBN: 9789027210104.

DOI: http://dx.doi.org/10.18290/rh.22708.8



Current Issues in Linguistic Theory 356

Language and Text
Data, models, information and applications

EDITED BY
Adam Pawłowski
Jan Mačutek
Sheila Embleton
George Mikros

JOHN BENJAMINS PUBLISHING COMPANY

The reviewed book *Language and Text. Data, models, information and applications* is the outcome of the 10th QUALICO 2018 conference of quantitative linguistics held for the first time in Poland (Wroclaw). QUALICO conferences have been organized under the auspices of International Quantitative Linguistics Association (IQLA) since 1991. The book is published by John Benjamins as the 356th volume of the series Current Issues in Linguistic Theory (CILT).

The theme of the conference "Information in language: Coding, extraction and applications" reflects the current increasing interest of scholars in the complex quantitative analysis of big data. Therefore, besides traditional quantitative linguistics disciplines such as stylometry and exploration of statistical laws of language, we can also see research analyzing big corpora using advanced deep learning techniques. This recent development of quantitative linguistics is reflected in the structure of the book. Thus, the book is divided into two sections: Theory and models and Empirical studies. A more traditional approach is presented in the first part, such as studying frequency distributions, their parameters, and their goodness of fit. In these studies, the main objective is to understand general patterns of using various language units such as words, syllables, syntactic functions. Part two of the book is devoted to studies that analyze big data and apply sophisticated advanced statistical methods. Both theoretical and practical studies can be found in the reviewed book.

The first section of the book is opened by a study written by R. Čech, P. Kosek and O. Navrátilová. In the chapter entitled "On the impact of the initial phrase length on the position of enclitics in Old Czech," the relationship between the length of the initial phrase and the positions of pronominal enclitics in a clause in Old Czech Bible translations is investigated. The authors hypothesize the negative correlation between the length of the phrase and the proportion of enclitics in the post-initial position.

Efficiency of two different collocation measures in corpus linguistics are tested in the chapter “Term distance, frequency and collocations.” One method is based on the frequency of terms while the other one is based on their distance. L. G. Johnsen shows on corpus of digitized texts from the Norwegian National Library (nearly 440,000 books) that both approaches have a small computational cost because they do not require computing a reference statistic.

V. Matlach, D. G. Krivochen, and J. Milička propose a new method for clustering texts in the chapter “A method for comparison of general sequences via type-token ratio.” The presented method provides a simple and efficient tool to describe the combinatorial behavior of sequences, vectorize them and use it for their comparison, clustering and estimating the quality of their generators. The method is based on type-token ratio and is so universal that can be used for comparison of any type of sequences without prior knowledge of their higher structures (random sequences, DNA, or source-codes).

The relation between syllable frequencies and syllable length in several Slavic languages are explored by B. Rujević, M. Kaplar, S. Kaplar, R. Stanković, I. Obradović, and J. Mačutek in chapter “Quantitative analysis of syllable properties in Croatian, Serbian, Russian, and Ukrainian.” The authors show that Zipf ’s law is not valid only for words but can be extended to syllables as well. Furthermore, they suggest a generalization of the Menzerath-Altmann law for a relation between the word length and the mean syllable length valid for both word types and word tokens.

H. Sanada deals with word order in Japanese in chapter “N-grams of grammatical functions and their significant order in the Japanese clause.” The author concludes that the time and the place appear between the subject and object with statistical significance. The occasion takes a position before the subject, between the subject and object, or after the object. Therefore, the occasion shows that Japanese is a free word order language. The subject and object play the role of ‘anchors’ in the clause.

P. Steiner is the author of the chapter “Linking the dependents: Quantitative-linguistic hypotheses on valency” where she studies a relation of syntactic and semantic aspects of case and valency to morphological properties. Steiner discovered that the larger the number of variables of a semantic predicate, the larger is the number of the syntactic dependents of the realized syntactic constructs, and the larger the number of semantic roles, the larger is the tendency for shortening on the syntactic level.

Relja Vulanović deals with grammar efficiency evaluation in chapter “Grammar efficiency and the One-Meaning–One-Form Principle.” The author compares various approaches for measuring grammar efficiency and proposes a new formula which is simpler and more straightforward to use compared to previous methods.

In chapter “Distribution and characteristics of commonly used words across different texts in Japanese,” M. Yamazaki uses a big corpus of modern Japanese texts of various genres to analyze frequency distributions of commonly used words. The author discovered several interesting findings. The distribution is similar to Zipf's law. Text length and the number of the texts do not affect the distributions. Additionally, as the text length increases, the number of commonly used words also increases linearly.

The second part, *Empirical studies*, starts with a theoretical chapter "The perils of Big Data." S. Embleton, D. Uritescu, and E. S. Wheeler discuss pros and cons of big data usage in quantitative linguistic. These days, large text databases are widely available and linguists use them regularly for research. They uncover many important new insights. However, big data and sophisticated tools can pose a number of potential pitfalls. The data should be carefully selected and separated into subsets so that patterns in the data are clear. It is also important to keep in mind the limits of applied methods.

M. Konca, A. Mehler, D. Baumartz, and W. Hemati deal with the distinctiveness of random and non-random texts based on text characteristics of quantitative linguistics in the chapter "From distinguishability to informativity: A quantitative text model for detecting random texts." The study shows that current random text models still generate texts that are easily distinguishable from non-random ones. Moreover, even small number of simple quantitative text characteristics is sufficient to show that.

G. Mikros and R. Voskaki present a newly developed tool for automatic readability text classification of Modern Greek texts according to the Common European Framework of Languages (A1 to C2) in the chapter "A Modern Greek readability tool: Development of evaluation methods." The tool is based on machine learning algorithm using several stylometric features used in quantitative linguistics. The accuracy of prediction readability level is 0.943 which is much more reliable than previous tools.

The relation of phoneme structure and success rate of Czech texts is investigated by J. Milička and A. H. Růžičková in their chapter on "Phonological properties as predictors of text success." The authors discovered that the popularity of online texts can be predicted by two phenomena (the beauty-in-averageness effect and the euphony principle).

M. Místecký analyzes texts produced by the candidates for the 2018 Czech presidential election by keywords and several traditional stylometric features such as lexical richness, thematic concentration or average word length in the chapter "Calculating the victory chances: A stylometric insight into the 2018 Czech presidential election." The results show that each candidate adopts a special strategy which can be tracked by the quantitative methods.

H. Moisl deals with visualization methods for non-linear high-dimensional data in the chapter "Topological mapping for visualization of high-dimensional historical linguistic data." The author proposes topological mapping, a non-linear visualization method coupled with a Self-Organizing Map, a specific topological mapping technique to plot and analyze typological characteristics of a small historic text corpus.

Automatic text classification is a common task in quantitative linguistics. In the chapter "The example of the bibliographic corpus of microtexts" written by A. Pawłowski, E. Herden, and T. Walkowiak use advanced statistical methods based on word embedding (word2vec, FastText) to develop supervised classification models applied the corpus of Polish microtexts, consisting of book titles. They reached very high accuracy of text classification where more than 70% of titles could be correctly assigned a writing species, while the accuracy of the gender recognition of the author was almost 80%.

A. Pawłowski, K. Topolski, and E. Herden compared data of Polish national bibliography from the period 1801–2019 to the general corpus of Polish language in the chapter "Quantitative analysis of bibliographic corpora: Statistical features, semantic profiles, word spectra." They show that these two corpora significantly differ in several features such as frequency distribution of parts of speech or lexemes.

The chapter "Analysis of English text genre classification based on dependency types" written by Yaqin Wang shows that syntactic features, namely dependency type, can be used as a distinctive text vector for classifying English genres. Several classification methods (principal component analysis, hierarchical clustering, and random forest) are applied. The results show that the biggest difference is between written and spoken texts.

A special chapter closes the book in memory of Gabriel Altmann, a pioneer of the quantitative approach in linguistics and IQLA's Honorary President since 2005, who passed away in March 2020. Several scholars share their memories of Altmann not only as an outstanding scientist and father of quantitative linguistics but as their teacher, colleague, and friend as well.

To summarize, the reviewed book *Language and Text. Data, models, information and applications* demonstrates the current trends in quantitative linguistics that are moving toward complex statistics methods applied to large datasets while still keeping interested in traditional fields (stylometry, analyzing distributions of various language units, finding general laws of language usage, etc.) at the same time. QUALICO conferences organized under the auspices of International Quantitative Linguistics Association (IQLA) provide a unique interdisciplinary forum for sharing quantitative approaches to linguistics. It is a unique opportunity for scholars from different research fields to exchange knowledge on quantitative approaches in linguistics. The diversity of methodological approaches and topics is also followed by a rich diversity of scholars from all over the world. Thus, many languages from different language families are analyzed, and no single language dominates. The reviewed book *Language and Text. Data, models, information and applications* is a wonderful source of inspiration and knowledge for scholars in the field of quantitative text analysis.

Miroslav Kubát, PhD
University of Ostrava
Department of Czech language
e-mail: miroslav.kubat@gmail.com
ORCID: https://orcid.org/0000-0002-3398-3125