Reviews

Peter Grzybek, Emmerich Kelih, Ján Mačutek (Eds.),
Text and Language: Structures · Functions · Interrelations. Quantitative Perspectives. Wien: Praesens Verlag, 2010. ISBN 978-3-7069-0625-8, 251pp.
Reviewed by **Haitao Liu** ((Zhejiang University, lhtzju@gmail.com)

This volume contains 23 presentations of the Qualico-2009, organized by IQLA (International Quantitative Linguistics Association). As revealed by the complicated title, these contributions explore structures, functions and interrelations of text and language from quantitative perspectives.

Following the short preface of the three editors is *Sergej Andreev's* quantitative analysis of Keats' style, which is concerned with the factors influencing genre differences.

Solomija Buk and others have attempted to spot word-length-related parameters in the Ukrainian language for the sake of automatic differentiation of text genres, arriving at the discovery that phoneme distribution can be a useful supplement to word-length-related parameters in genre classification.

Radek Čech and Ján Mačutek have examined the distribution of valency frames in Czech and tested the hypothesis of a relationship between word length and the number of valency frames Their work shows that the distribution of valency frames well fits the Good distribution and "the shorter the verb, the more valency frames ".

Łukasz Debowski announces recent developments of a new probabilistic explanation of the Zipf law and the Herdan law, and reports, in particular, his work in correlating the number of set phrases in a text and the number of described facts.

On the basis of 120 Slovenian texts of four different genres, *Gordana Đuraš* and *Ernst Stadlober* have modelled word length frequencies according to the Singh-Poisson distribution.

With their experience in the application of Multidimensional Scaling (MDS) to geolinguistic data, *Sheila Embleton* and others investigate approaches to testing quantitative hypotheses.

Peter Grzybek's article *covers* the measurement of text difficulty. He has found that, at least for German, text difficulty can be measured without any language-specific adaptation; and probably no text typological specifics need to be considered when Tuldava's TD is employed to measure text difficulty.

Emmerich Kelih presents empirical Serbian evidences of the Menzerath law and advances linguistic interpretations of the usually iteratively determined parameters which can be replaced by the mean syllable length of one-syllable words.

Based on textual properties which are purely formal and automatically determinable, *Reinhard Köhler* and *Sven Naumann* propose a syntagmatic approach to automatic text classification, which facilitates not only automatic classification of text but also the linguistic understanding of text properties.

Jan Králík probabilistically explains Zipf's law. According to the author, the utility hierarchy, which is introduced in his paper, can accout for the working mechanism of the prin-

ciple of least effort.

Jerónimo Leal and *Giulio Maspero* quantitatively probes into Tertullian's authorship of the *Passio Perpetuae,* showing the advantage of entropic distance over bigram distance in classifying the texts studied in their research.

Sylvain Loiseau explores the hypothesis that the plurality of axes of variation may be useful in textual typology. This paper introduces some variationist viewpoints into text typology and corpus-based analyses of variation.

Ján Mačutek illustrates, with an example whose data well fits the chi square goodness of fit test, the ways to avoid the pitfall that may occur in the investigation of rank-frequency distributions.

Gregory Martynenko argues that a Weibull-approximation of the empirical dependencies "vocabulary size - sample size" is useful in both establishing hypothetical borders of the lexical diversity and seeking a rule for the harmonious organization in vocabulary size.

Ivan Obradovic and his colleagues contribute two papers to this volume. The first paper, which focuses on wordnets as a means for refining queries in IR tasks, advances a set of simple and natural relevance indices for tuning the query formulation process. The second paper, which is concerned with the distribution of canonical syllable types in Serbian, points to, in spite of the negative results, some interesting directions worthy of further investigation.

Vasilij V. Poddubnyj and *Anastasija S. Kravcova,* with their investigation into statistical reduction of the feature space of text styles, argues that the transformation of the feature space is helpful to find a minimal set of statistically independent latent features.

Andrij Rovenchak and *Valentin Vydrin* quantitatively explores properties of Nko, an indigenous writing system of Manding languages of West Africa, including script's complexities, their interrelations with frequencies, the grapheme-phoneme correspondence and phoneme distribution.

Haruko Sanada's exploration into the distribution of motifs and the relationship between length and frequency of motifs in Japanese texts registers that motifs, which are defined as word-length sequences following time-series, are correct conceptual abstractions subject to the same laws as all other linguistic units.

Tatiana Sherstinova reports the quantitative data processing in the ORD speech corpus of Russian everyday communication. This kind of corpus benefits both the investigations into Russian communication strategies and the descriptions of the vocabulary and grammar of modern spoken Russian.

O.G. Shevelyov and *V.V. Poddubnyj* present how to conduct complex investingations of texts with "StyleAnalyzer", a software tool enabling researchers to carry out various investigations in the fields of quantitative and computational linguistics. With the objective of creating a dictionary of Japanese collocations,

Tadaharu Tanomura discusses several issues of the retrieval of collocational information from Japanese corpora.

Nicolas Turenne examines the influence of time on the distributions of both content-word occurrences (or named entities) in texts and clusters of content-words where these words are linked due to sharing contexts.

At the end of this volume are listed, in addition to subject and author indexes, addresses of all authors — a thoughtful arrangement facilitating not only retrieving needed information

in this volume but contacting the authors concerned.

In summary, this volume includes both the topics of purely traditional quantitative linguistics and contents of computational linguistics, corpus linguistics, literary stylistics, etc. Even those papers concerning topics of purely quantitative linguistics are somewhat oriented toward practical applications such as text classification. Perhaps, this trend reflects that quantitative linguistics involves in fact interdisciplinary research which can assist significantly practical studies such as text processing. This volume is very useful for readers to understand the latest developments of quantitative linguistics.