There has been numerous publications concerning quantitative linguistics, five of which are proceedings of the International Conference on Quantitative Linguistics (QUALICO) held in 1991, 1994, 2009, 2010 and 2012 respectively. QUALICO was held for the first time in Trier, Germany in the year 1991 as an international conference to demonstrate the state of the art in quantitative linguistics. And the International Quantitative Linguistics Association (IQLA) was founded during the second conference in Moscow in 1994. Then the follow-up conferences have been organized regularly since 1997 by IQLA (QUALICO 1997 in Helsinki, Finland, QUALICO 2000 in Prague, Czech Republic; QUALICO 2003 Athens, Georgia USA, 5th Trier Symposium on QL2007 and QUALICO 2009 in Graz, Austria).

This book mainly focuses on the methods and applications of Quantitative Linguistics based on the 16 selected and peer-reviewed papers of the 8th QUALICO held in Belgrade, Serbia during April 26–29, 2012. Its editors, Ivan Obradović, Emmerich Kelih, Reinhard Köhler have classified the topics delivered during the conference into four categories: stylometrics, Fundamental linguistic research, application-oriented research, and methodological issues.

First of all, two papers are categorized under the problem of stylometrics, addressing the authorship attribution or text classification in relation to stylistic features revealing the author’s membership in social or other groups.

Dario et al. combine the traditional philological approach (based on N-grams) and the recent mathematical technique (based on data compression and its role in the estimation of the entropy of a source text) to a real case in Authorship Attribution – the attribution of a letter written in Greek by two brothers. The final result reveals that Ep. 38 is attributed to Gregory with a high precision of 97% in the cases of complete agreement of all the four methods designed under the above two main methods. This should be the highest degree of precision that people have ever achieved in authorship attribution. Moreover, they also introduce a novel method to analyze the texts of the corpus at small scale. This method

Aiyun Wei: College of Foreign Studies, Guangxi Normal University, No. 15 Yucai Road, Guilin, CHN-541004, China. E-mail: wendyaiyun@163.com
tries to compute an index which can indicate the intersection between two regions written by different authors. The results prove that the method is very useful in general cases of authorship identification and it can show an internal structure of the work.

Mikros’s investigation is dedicated to methods for authorship attribution and gender identification in a corpus of blogs written in Modern Greek language. The results show that using the stylometric features and the state-of-the-art machine learning algorithm (SVM), the accuracy in authorship attribution and gender identification is 85.4% and 82.6% respectively. This study also yields some important findings: specific character n-grams carry significant authorship information while specific word n-grams have increased importance in author gender identification and word length measures convey both authorship and gender evidence; author identification and gender detection are two different tasks with distinct patterns of stylometric feature interaction; and author gender is conveyed through specific syntactical and morphological patterns while authorship seems to rely on over-or under-representation of specific high frequency words.

Secondly, the category of fundamental linguistic research includes six papers.

Beliankou, Köhler, and Naumann attempt a quantitative investigation of argumentation structures based on Potsdam Commentary Corpus (German). They apply a method based on the analysis of motifs for linguistic units and properties such as word length, frequency and polytextuality, and the results show that this method can be used to analyze RSR structures, theoretical probability distributions can be found as models of the text phenomena discussed, and these mathematical models can also be linguistically interpreted.

Peter Grzybek approaches the close and distant relatives of the sentence in Russian texts from a Menzerathian perspective, focusing on the relation of the sentence to linguistic entities from “indirect” neighbors: the sentence length ↔ word relation (skipping the intermediate level of clauses) length and the sentence length ↔ chapter length relation (skipping the intermediate level of paragraph). The results show that the Menzerathian principle, which models the constituent relation of linguistic entities from two directly neighboring levels, is also at work in such ‘indirect’ neighbors. Peter Grzybek also puts forward that such results must be extended to other languages other than Russian and future studies should have to take into account possible text-type specifics besides literary texts. This is the first successful attempt to prove such cross-level relations and it will shed important insight on general principles of global text processes across levels of different components of texts.

The diversification of the cases of Czech nouns is discussed in Ján Mačutek and Radek Čech’s paper, in which the relationship among frequency, morphology
and phonetics is observed. The study employs two approaches: one is nominative singular as the basic form of a noun regardless of its grammatical number and the other is nominative singular as the basic form of a noun in the singular and nominative plural as the basic form of a noun in the plural. The findings indicate that in both cases, the declensional morphology of Czech nouns is clearly related to the frequency, or to be more specific, “the lesser the change, the higher the frequency”, which satisfies the Principle of the Least Effort derived from previous studies and also sheds new light on the investigation of morphology. And further development of the study could be seen in two directions: similar studies for other languages (other than Slavic) with rich flexional system and other parts of speech, adjectives in particular.

Vasiliy Poddubnyy & Anatoliy Polikarpov’s paper is a further investigation of the series of studies in creation and experimental verification of a novel mathematical model of language evolution, namely, stochastic dynamic model of evolution of language sign ensembles. The paper explores in details all the procedures of the special model, from the major linguistic preliminaries, the initial formalization of the main points, the mathematical modeling of sign life cycle and that of the ensemble behavior of signs in language, the testing of the model for polysemy distribution of lexical signs, the interpretation of polysemy data to interpretation of frequency data. The results show that this model follows the Principle of the Least Effort, assumes a Poisson birth flow for language signs and an exponential distribution for their associative-semantic potential (ASP) and also enables us to deduce equation for synchronous lexical polysemy and rank-frequency distributions.

Matilde Trevisani and Arjuna Tuzzi try to unveil chronological patterns in the frequency of words in a text corpus, attempting to identify a specific sequential pattern for each word as a functional object (as a curve) and group these curves in order to group words with similar temporal patterns. The corpus analyzed in the study includes 63 end-of-year messages delivered by ten presidents of the Italian Republic over the period from 1949 to 2011. To overcome some of the limitations of similar studies based on simple form-types, they based their study on lemmas, processed the textual data by means of Taltac2 dedicated software, carried out the statistical analysis with R (curvclust) and found that many clusters are representative of a certain president while other clusters also show trends associated with events and history. This kind of functional data analysis is a growing area in statistics and proved promising but is still waiting for justifiable linguistic theories.

Makoto Yamazaki tries to capture the characteristics and structure of text through a quantitative analysis of lexical cohesion, using the co-occurrence rate to calculate the degree of cohesion based on the DVD edition of the Balanced
Corpus of Contemporary Written Japanese (BCCWJ), which includes a variety of individual texts such as Newspapers, Magazines, White Papers, School Textbooks, Minutes from the National Diet and etc.. The results suggest that the degrees of cohesion differ among different types of texts, with high degrees of cohesion in laws, white papers and minutes from the National Diet and low degrees of cohesion in newspapers, bestselling books and magazines. The study also finds out the reason for the low degree of cohesion in literature and that noting changes in the co-occurrence rate could be a way to examine the segmentation of texts.

Thirdly, another group of three selected papers address the application-oriented issues of quantitative linguistics.

Gordana Đuraš, Ernst Stadlober and Emmerich Kelih analyze the distribution of word length in 120 Slovenian and 120 Russian texts of four different types – journalism, poems, private letters and prose. They use the index of dispersion $d = \frac{s^2}{\bar{x} - 1}$ as a classifier of texts and suggest two-parametric generalizations of the Poisson distribution to deal with the over-and under-dispersion. The results reveal that the mean word lengths in Russian texts are significantly longer than those in Slovenian texts and the linguistic explanation for the phenomenon is also given. They also apply different estimation procedures such as the method of moments, the maximum likelihood method to their Slovenian and Russian texts. Even though Slovenian and Russian are rather different on all linguistic levels, their study has proved that the word length frequency distribution of the sampled texts can be described by one theoretical model – the Singh-Poisson model with two-parametric generalization of the Poisson distribution.

Ivan Obradović discusses a method to extract translational equivalents (semantically related word pairs) from aligned (at sentence level) texts of a Serbian-English corpus which is related to education, finance, health and law. He lemmatized the corpus, applied the method on lemmas of word forms from the corpus and extracted candidate translational equivalents relying on a simple ranking method based on word co-occurrence. The method was also validated on the SELFEH corpus, choosing Serbian for the source and English for the target language. The study has proved that the method succeeded in offering an acceptable translational equivalent for a little less than 80% of the SELFEH corpus.

Haruko Sanada uses Altmann’s and Popescu’s Thematic Concentration (TC) as an indicator to study the characteristics of 10 Japanese prose texts and finds that TC has a relation to the number of different words in a text, but has none relation to the length of a text. A modified TC is also proposed and its $u$ test is used to investigate how Japanese prose texts are similar from the point of view of TC. The results indicate that the degree of TC depends on the text type, namely, scien-
Scientific texts have greater TC and art, social sciences poetry follow it while prose has relatively smaller TC.

Finally, the last category of papers concentrates on discussions of methodological topics.

Łukasz Dębowski devotes his study to the specific topic of verifying Hilberg’s conjecture that information between two adjacent text blocks of length \( n \) is proportional to \( n^\beta \) (\( n \) is the block length). The test is conducted for a selection of English prose using the Lempel-Ziv algorithm, yielding a result not excluding Hilberg’s conjecture with an exponent \( \beta \) close to 1, but proving an exact upper bound for the exponent to be 0.949. This is the first time precise calculation of the conjecture and may mainly be of interest from the perspectives of mathematics and informatics.

Different string similarity methods for cognate identification are assessed by Antonella Delmestri and Liviu P. Dinu. They focus their study on different static schemes, including Manhattan Distance, Jaro and Jaro/Winkler Distances and ALINE, and several learning methods, such as Pair Hidden Markov Models, Dynamic Bayesian Networks and a measuring string similarity system. They first introduce those manually designed procedures and data-driven methods, and then compare the results achieved by the manually-designed procedures and those by the data-driven models. The outcomes suggest that all the orthographic learning methods may accurately detect traces of sound changes left in the orthography and perform better than the static methods, even though the static methods may have been specifically designed for phonetic alignment.

Jiří Milička investigates the mathematical relation between TTR (type-token relation) and rank-frequency distribution, presenting different formulae to exactly compute TTR, \( T_1 \text{TR} \) (hapax legomena-token relation) and \( T_g \text{TR} \) (the relation between the number of types of certain frequency and the number of tokens) from a distribution of types from a frequency distribution of types (or RFR) and from \( T_1 \text{TR} \), while the calculation of \( T_g \text{TR} \) from TTR still need an approximation. The study also observes that, for a very large corpora, the ratio between the number of hapax legomena and types converges to a constant \( Z \) (\( Z > 0 \)), and under this assumption, it proves that for an infinitely increasing number of tokens, the number of types also increases unlimitedly.

Focusing on the measurement of nonlinear distance in data derived from linguistic corpora, Herman Moisl first defines the nature of nonlinearity and describes two different existing ways to detect nonlinearity, and then he introduces the ratio of mean graph to mean Euclidean distance between all pairs of nodes in a graph to measure the amount of nonlinearity in a data manifold. And finally, he uses a case study based on data abstracted from the Diachronic Electronic Corpus of Tyneside English, a dialect and verifies all the aforesaid methods with
the data, which shows that all the methods agreed that substantial nonlinearity did occur.

By generalizing the logistic differential equation to a system of n first-order partial differential equations, Relja Vulanović extends the Piotrowski-Altmann Law to several dimensions. The solution to the system is a multidimensional generalization of the sigmoid which is an S-shaped curve in 2 dimensions and is used earlier to model certain phenomenon about parts-of-speech systems, while this study applies multi-dimensional variants of the law based on data collected from those previous studies, so that with its derivation, explanation, and application, the Multidimensional Piotrowski-Altmann Law is fully established.

The selected proceedings provide a broad overview on current research topics of quantitative linguistics and related areas and in each study, the prospects of future research concerns of related issues are also covered. As a whole, this collection presents the following features:

1. It includes a wide range of issues, concerning a lot of aspects of quantitative linguistics – gathering the issues under four main topics as aforesaid, each topic dealing with a specific issue, ranging from verifying existing theories, applying existing methods in a new field, extending proved law to some new dimensions, looking into attributes or properties of particular texts, to probing into new methods for unsolved problems.

2. It covers the researches on the corpora collected from quite a number of languages including English, Modern Greek, German, Russian, Czech, Italian, Japanese, Slovenian, Serbian and Tyneside English, which may open encourage using methods investigated for further studies of different language or even large variety of dialects.

3. It investigates the corpora of a large variety of text styles, involving letters, blogs, prose, dictionaries, magazines, newspapers, argumentative texts, full-length novels, bestselling books, library books, school textbooks, Q&A bulletin boards, presidential end-of-year speeches and etc, which may broadly enrich the applicability of quantitative linguistics.

With these features, the collection obviously serves as an important reference for both beginners and respected researchers in the field of quantitative linguistics.