Journal of Quantitative Linguistics
2012, Volume 19, Number 2, pp. 162–170
http://dx.doi.org/10.1080/09296174.2012.659002

Routledge
Taylor & Francis Group

BOOK REVIEW

Grzybek, Peter, Kelih, Emmerich and Mačutek, Jan (Eds). *Text and Language*. Wien: Praesens Verlag, 2010. ISBN 978-3-7069-0625-8 (pbk), VIII + 251 pp.

Reviewed by Fan Fengxiang
School of Foreign Languages, Dalian Maritime University

Since the first quantitative linguistics conference held at the University of Trier, Germany, in 1991, five more QUALICO's have taken place in various places both in the New and Old Worlds on a wide spectrum of topics concerning "the multitude of quantitative properties which are essential for the description and understanding of the development and the functioning of linguistic systems and their components" (Köhler & Altmann, 2005, p. 12). These topics cover all the 10 major areas of quantitative linguistics summarized by Köhler, Altmann and Piotrowski (2005) and more, since quantitative linguistics now "has grown to a degree which makes it difficult to maintain an overview over the many topics and objects of investigation, the models and methods applied and developed, and the various results published in books and in several journals" (Grzybek & Köhler, 2007, p. viii).

The latest QUALICO was held in Graz, Austria, in 2009. At the conference, 23 papers by scholars of international status were presented, which were subsequently put into a book entitled *Text and Language* with a subtitle *Structures, Functions, Interrelations, Quantitative Perspectives*, published by Praesens Verlag in 2010, with Grzybek, Kelih, and Mačutek as editors. As the title suggests, the 23 contributions of the book mainly focus on the quantitative analyses and measurement of text components, constructs, functions and distributions involving 12 different languages. In addition, topics on some well-known linguistic laws, problems with statistical methods and applications of corpora and software packages for language research

are also dealt with. The topics of these articles can be tentatively grouped into the following areas.

1. Text categorization and authorship attribution. There are six contributions on this area.

''Quantitative analysis of Keats' style: genre differences'' by Sergej Andreev. The author uses discriminant analysis to study the genre styles between Keats' sonnets and odes. Unlike many other similar studies, the author makes distinctions in selecting data sources according to their popularity rate. In the discriminate model, the following features are included: inexact rhyme, omission of stress on ictuses 1, 2, 3, 4, 5, feminine and dactylic clausula, and emphatic end. The result shows there exist significant stylistic differences between Keats' sonnets and odes.

''Word-length-related parameters of text genres in the Ukrainian language, a pilot study'' by Solomija Buk, Olha Humenchyk, Lilija Mal'tseva and Andrij Rovenchak. In this study, the authors use word-length related parameters to study genre differences in Ukrainian. These parameters include mean word length in syllables, dispersion of word length and phonemes, etc. The advantage of using these parameters is that these parameters can be automatically obtained using raw texts, i.e. untagged texts.

''A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics'' by Reinhard Köhler and Sven Naumann. Generally, methods used in text classification are based on distribution models of word frequency, word length, sentence length, utilization of specialized vocabulary, etc. Some use syntactic information for such purposes. But the authors employ motifs, which are a continuous series of equal or increasing values such as length or frequency of linguistic elements-words, phrases, sentences etc. The result demonstrates that using motifs as the unit of analyses in text classification is vocabulary independent and involves fewer attributes.

''Revisiting Tertullian's authorship of the Passio Perpetuae through quantitative analysis'' by Jerónimo Leal and Giulio Maspero. The authors employ two measures, the entropic distance and the bigram distance for the attribution of the Introduction and Conclusion of *the Passio Perpetuae*. Two sets of samples are used; one consists of works by Tertullian, the other of non-Tertullian works. The conclusion supports the attribution of the Introduction to Tertullian and that the entropic distance is more effective in attribution studies of this kind.

"Textual typology and interactions between axes of variation" by Sylvain Loiseau. In this article the author argues that better textual typology can be achieved by taking into account multi-axis of variation such as time, author, genre and section within a text. Discriminant morpho-syntactic features on a particular axis are influenced by other axes.

"Statistical reduction of the feature space of text styles" by Vasilij V. Poddubnyj, and Anastasija S. Kravcova. The texts style of different writers can be characterized by different sets of features; some are more discriminatory while others are less so. In addition there are "noise" features. The authors adopt both principle component analysis and discriminant analysis for text style characterization. 80 large works of fiction by 11 Russian writers of the 19th century are used in the study. The result shows that discriminant analysis provides better discrimination of authors.

2. Distribution and measurement of linguistic units in texts. Six contributions are on this area.

"On the quantitative analysis of verb valency in Czech" by Radek Čech and Ján Mačutek. This article examines the distribution of valency frames in Czech and tests the hypothesis that the shorter the verb, the more the verb valency frames. The result shows the distribution of valency frames is regular and can be fitted with the Good distribution. The hypothesis of the relationship between number of valency frames and word length is also corroborated.

"Distribution of canonical syllable types in Serbian" by Ivan Obradović, Aljoša Obuljen, Duško Vitas, Cvetana Krstev and Vanja Radulović. In this article the authors examine the distribution of canonical syllable types in Serbian. They first use the Zörnig-Altmann model which successfully describes the canonical syllable type distribution in a sample in Indonesian. The model employs the discrete two-dimensional approach to the application of a truncated Conway-Maxwell-Poisson distribution. However, this model fails to achieve the intended target. They also use the two-dimensional negative binomial distribution adopted by Beőthy and Altmann for semantic diversification of Hungarian verbal prefixes. Again this attempt is unsuccessful. This shows that the distributions of certain linguistic constructs are language specific.

"Distribution of motifs in Japanese texts" by Haruko Sanada. In this contribution, the author mainly examines the frequency spectrum of motifs and the relationship between the length and frequency of motifs in Japanese. The results show that both of them are regular; the former can

be adequately captured by a mathematic model from Popescu et al., and the latter by a model derived from Wimmer, Altmann and Köhler. The author poses some questions as to the application of motifs to other postpositional languages and whether motifs are "legal" linguistic units or very high abstractions. However, in the case of Japanese, motifs are "correct" conceptual abstractions abiding by the same laws just like any other linguistic units.

"Modeling word length frequencies by the Singh-Poisson distribution" by Gordana Đuraš and Ernest Stadlober. According to the authors, in modelling word length distributions the negative binomial distributions seem to be adequate for count data with over-dispersion, i.e. $d > 1$, $d = s^2/(\bar{X} - 1)$; as for under-dispersed count data, i.e. $d < 1$, Dacey-Poisson models are suitable. This article proposes the Singh-Poisson for covering the whole range of d: $d > 1$, $d = 1$ and $d < 1$. This model is tested on 120 Slovenian texts, and reasonable and stable parameters of the model α and θ are obtained and a good fit is achieved.

"Measuring lexical richness and its harmony" by Gregory Martynenko. In this contribution the author uses the Weibull function to model the vocabulary growth of three Russian writers at the beginning of the twentieth century. The asymptotes of the vocabulary sizes of the three writers are given and extrapolations of the vocabulary growth approaching the asymptotes are made. The principles of the golden section, i.e. multiplicativity and additivity are applied in order to outline hypothetical borders of the lexical diversity in Russian prose at the beginning of the twentieth century.

"Measuring semantic relevance of words in synsets" by Invanić Obradović, Cvetana Krstev and Duško Vitas. This contribution is on the optimization of queries for information retrieval from Serbian texts. To establish an optimal balance between recall and precision in an information retrieval system, two natural relevance indices for measuring semantic relevance of words in synsets are developed using the Serbian Wordnet. These indices are tested on a small sample of selected words and corpora and the results prove the usefulness of the indices.

3. Interpretation of some linguistic laws. Four of the contributions explore this area.

"A link between the number of set phrases in a text and the number of described facts" by Łukasz Dębowski. This article gives a new probabilistic interpretation of the Zipf and the Herdan law. That is, if an n-letter

long text describes n^{β} independent facts in a repetitive way, where $0 < \beta < 1$, then the text contains at least $n^{\beta}/\log n$ different set phrases.

"Text difficulty and the Arens-Altmann law" by Peter Grzybek. This article deals with measures for text difficulty. There are over 200 published formulae for measuring text difficulty, but the Flesch Reading Ease Index (REI) is the most popular; there is a less well-known, but very important formula for measuring text difficulty (TD), proposed by Tuldava. TD and REI have a very high correlation for German texts, hence measuring text difficulty using TD seems to be a very promising way. The controlling mechanism behind REI and TD can be explained within synergetic linguistics and the Arens-Altmann law and, since the latter states the relationship between word length and sentence length, perhaps both REI and TD can be reduced to only one variable, using either word length or sentence length.

"Probabilistic reading of Zipf" by Jan Králík. This article offers a new interpretation of rank in the Zipf law. The author theorizes that rank is closely connected with the corresponding measure of utility; the latter can be expressed as the relative frequency or probability of a word. The significance of the new interpretation lies in that the construction of each parameter preserves real meaning so that parameters can be analysed, measured and interpreted.

"Parameter interpretation of the Menzerath law: evidence from Serbian" by Emmerich Kelih. This contribution uses Serbian texts to test the Menzerath law, i.e. $SyL = a \cdot WoL^{-b}$ where SyL is the syllable length measured in number of graphemes and WoL the word length measured in number of syllables. The result shows a very high R^2. a is commonly understood to be the starting value of the fitting curve while b the steepness of the decrease of the curve. However, according to Köhler a actually represents the mean length of a language construct consisting of one component, here meaning the mean syllable length of one-syllable words. Replacing a with the mean syllable length of one-syllable words still achieves a satisfactory R^2 ($R^2 > 0.92$). As proposed by Köhler, a and b have a linear interrelationship. Here replacing a with the mean syllable length of one-syllable words still show a linear tendency.

4. Corpus and software package application in linguistic research. Four contributions are on this area.

"Retrieving collocational information from Japanese corpora: its methods and the notion of 'circumcollocate'" by Tadaharu Tanomura.

The author discusses methods for retrieving collocational information from a Japanese corpus built by the author from web pages. Extracting collocational information from Japanese texts is quite challenging because there are no white-space word demarcations in Japanese texts. A software package MeCab is used to identify words. The most effective method for collocational information retrieval is looking at the co-occurring *N*-grams on the word level.

''Diachrony of noun-phrases in specialized corpora'' by Nicolas Turenne. This article examines the impact of time on the distribution of content words and content word clusters in specialized corpora. Analyses show that the distribution of content words does not correspond to already known laws discovered about diachronic phenomena. The size of associations of content words can be a mixed distribution of small and double size associations.

''Quantitative data processing in the ORD speech corpus of Russian everyday communication'' by Tatiana Sherstinova. This article introduces the ORD speech corpus. The corpus contains diverse genres and styles of speech made by demographically balanced group of 40 people. At the time of writing, the corpus contains 994 communicative episodes. The corpus is annotated on multiple linguistic and paralinguistic levels such as transcript of speech, speaker's code, words, morphemes transcribed in IPA, grammatical type of morphemes, spelling of morphemes, quality of speaker's voice, comment on phrases, etc. General statistics of the corpus are obtained on frequency lists of sentences, syntagmas, auxiliary annotation symbols, word forms, etc. Corpora of this kind are extremely useful in fundamental linguistic research and in natural language processing.

''Complex investigation of texts with the system 'StyleAnalyzer'' by O. G. Shevelyov and V. V. Poddubnyj. This article introduces the software package StyleAnalyzer that can perform complex text analysis. The software was developed at the computer science department of Tomsk State University. A new generation StyleAnalyser is being planned.

5. Problems concerning results of statistical tests. Two contributions are on this area.

''How do I know if I am right? Checking quantitative hypotheses'' by Sheila Embleton, Dorin Uritescu and Eric S. Wheeler. In this article, the authors raise questions as to the validity of results obtained from statistical tests. They use Multidimensional Scaling to study dialect

variation in the North-West Region of Romania using the entire set of data; however, the resulting dialect picture is not close to the geographic map. Then the authors use a collection of interpretive maps which reflect the judgment of trained scholars, and the result corresponds to the dialect picture, and this is more in line with the expected analysis of the data. The authors argue that although every piece of data is important, an educated and well-considered selection of data is more valuable.

"Rank-frequency distributions: a pitfall to be avoided" by Ján Mačutek. In this article, the author shows that the result of a statistic test should not be accepted blindly. If a statistical test rejects a hypothesis it should be taken as a suggestion, not a final verdict. If a theoretical model fits well to empirical data it does not mean the model and the observed distribution are the same; it only means they are not far apart. To demonstrate this point, the author uses modified rank-frequency data from a poem by J. W. von Goethe. The right truncated zeta distribution achieves a very good fit ($\alpha = 0.5202$, $n = 225$, sample size = 326). A chi square goodness of fit results in a p-value = 3. Then the author uses the right truncated zeta distribution model with the same α, n and sample size to generate 100 random samples, thus 100 means and variances are obtained. The mean of the 100 means and the mean of the 100 variances are compared with the empirical data using the t-test. Counter to intuition, there exists a significant difference between this mean and the empirical mean.

6. Script analysis.

There is one contribution on the script analysis of the Nko writing system. The paper "Quantitative properties of the Nko writing system" by Andrij Rovenchak and Valentin Vydrin studies the quantitative characteristics of the Nko writing system, which was created in 1949 for the Manding languages in West Africa. It describes the phonological system of Maninka-Mori, a Manding variant and the phoneme-grapheme relationship, obtaining the mean orthographic uncertainty and calculating the script complexity.

Of these contributions in the book, two articles, "How do I know if I am right? Checking quantitative hypotheses" and "Rank-frequency distributions: a pitfall to be avoided" strike a different note among the other statistic-related and data-intensive articles. The authors argue for exercise of caution towards statistical data and test results instead of applications of certain statistical methods and techniques. Statistics, according to Gomez (2002), allows linguists to draw inferences from

complex numerical linguistic data while intuition cannot always reveal something significant about linguistic variation. However, when statistical results are contrary to common expectation, both data and the statistical method used should be thoroughly rechecked. A good case in point is the use of the one-sample Kolmogorov-Smirnov Test for normality. It is particularly sensitive to the size of data. A normally distributed set of data would often end up as otherwise simply because the set of data is too large. In this case, we need to switch to other tests available instead of accepting the test result.

As we have already seen, one article, ''Diachrony of noun-phrases in specialized corpora'', tackles its object of study diachronically. This shows that the quantitative approach is very useful and promising in historical linguistics and the study of language change. A number of researchers have employed this approach in these fields. For example, Lieberman et al. (2007) studied the relationship between frequency and lexical change and found that in the evolution of the English language, high frequency irregular verbs, such as *come*, *go*, *get*, etc., retain their irregularity while those of lower frequency are regularized. Atkinson (2011) examined the number of phonemes of 504 languages in the world to trace language origin and the results indicate older languages tend to have more phonemes, and the conclusion is that all human languages originated in Africa.

On the whole, the book deals with the complex relationships quantitatively among the linguistic constructs and components, both concrete and abstract, within a text or sets of texts in a language. As the editors of the book say in the preface, these contributions involve mathematics, statistics, information science, computer linguistics, corpus linguistics and literary scholarship, apart from traditional linguistics, which reflect the interdisciplinary nature of quantitative linguistics. And this makes the book a valuable food for thought for students and researchers in these areas.

REFERENCES

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science, 332*(6027), 346–349.

Gomez, P. C. (2002). Do we need statistics when we have linguistics? *DELTA, 18*(2), 233–271.

Grzybek, P., & Köhler, R. (2007). Viribus quantitatis. In P. Grzybek and R. Köhler (Eds), *Exact Methods in the Study of Language and Text*. (pp. vii–xiii). Berlin: Mouton de Gruyter.

Köhler, R., & Altmann, G. (2005). Aims and methods of quantitative linguistics. In G. Altmann, V. Levickij, & V. Perebyinis (Eds), *Problems of Quantitative Linguistics: A Collection of Papers*. (pp. 12–41). Chernivtsi: Ruta.

Köhler, R., Altmann, G., Piotrowski, R. (2005). Preface. In R. Köhler, G. Altmann, & R. G. Piotrowski (Eds.), *Quantitative Linguistics, an International Handbook*. (pp. viii–x). Berlin: Walter de Gruyter.

Lieberman, E., Michel, J., Jackson, J.,Tang, T., & Nowak, A. M., (2007). Quantifying the evolutionary dynamics of language. *Nature, 449*(11), 717–721.