

Book Review

Emmerich Kelih, Róisín Knight, Ján Mačutek, and Andrew Wilson (eds.): *Issues in Quantitative Linguistics Vol. 4. Dedicated to Reinhard Köhler on the Occasion of his 65th Birthday*. Lüdenscheid: RAM-Verlag, 2016. 288 pp. ISBN: 978-3-942303-44-6.

Reviewed by **Ruina Chen**, School of Foreign Languages, Guizhou University, Guiyang, China, E-mail: chenruina2012@qq.com

<https://doi.org/10.1515/glot-2017-0008>

The fourth volume of this specialized series has been dedicated to one of the fathers of quantitative linguistics in Germany. R. Köhler is the initiator of linguistic synergetics, elaborator of background theories, introducer of motifs, specialist for computers and informatics, founder of the *Journal of Quantitative Linguistics*, founder of the International Quantitative Linguistics Association, etc. It is not easy to enumerate all of his merits for this discipline. In intimate circles one calls him the “King of QL”.

It is not necessary to present his bibliography, it would take too many pages. The editors of the volume belong to three nations and the contributors to 13 different ones. There are linguists, mathematicians, physicists, information and computer scientists. The volume begins and ends with Altmann’s humorous preface and a gag about retired professors (286–287). The other articles are serious.

The volume begins with the last discovery of Köhler, namely “motifs”, taken from musicology and representing today a quite usual linguistic unit (Altmann, G., *On Köhlerian Motifs*, 2–8). Liu, H. and Fang, Y. (*Quantitative aspects of hierarchical motifs*, 9–26) show that starting from the dependency grammar it is possible to define also hierarchical motifs. They apply the results to Czech, Chinese and English. Milička, J. (*Key Length Motifs in Czech and Arabic Texts*, 27–42) analyses length-motifs in Czech and Arabic and proposes a new type of scaling. Evidently, this new unit seems to be a good way to create a more abstract entity based on sequential and hierarchic data. Needless to say, motifs can be found also in any type of script. Thus the discovery of motifs seems to have the same scientific value as the discovery of phonemes many years ago.

Language complexity is, as a matter of fact, a problem that can be described in many ways. First, the definitions of complexity may strongly differ; second, the relations between phenomena creating it may be conjectured differently; third, one may ask how many languages are necessary to give a satisfactory answer. G. Colomá in *A Synergetic Regression Model of Language Complexity Trade-Offs* (41–60) takes into account phonology, morphology, syntax and

lexicon and searches for their relations in 100 languages, an enormous performance that will surely find a number of followers especially under typologists.

A full-fledged synergetic study based on Köhler's system can be found in Lu Wang's *Synergetic Studies on Chinese Lexical Structure* (61–81). The properties: frequency, word length, polyfunctionality, polytextuality, polysemy and their interrelations as applied to Chinese data are presented. All formulas and computing results are displayed. Besides, the Köhlerian philosophy of linguistic theory formation is reviewed. It is good to know that a linguistic theory functions and penetrates slowly into non-IndoEuropean languages.

The study of the distribution of parts of speech in Japanese texts is the object of the article *A Measurement of Parts of Speech in Texts Using the Noun-Based Proportion* by H. Sanada (82–93). The problem has a long history in Japan where the research concerning it began about 60 years ago. Well known are Ohno's, Mizutani's, and Kabashima's contributions which are the basis of this article. The author constructed only 4 classes of POS (N = noun, V = verb, A = adjective, adverb, I = conjunction, interjection) the other ones are collected in a class called O. The results of cumulative distributions are presented both numerically, formally and graphically. The article opens a whole field of possible investigations analogous to the study of the Busemann-Altman indicator in Europe. Texts may be characterized by indicators or linear functions, they may be further classified. It would be interesting in general what kind of functions one obtains if one classifies the POS in a different way.

The article *Testing Hypotheses on English Compounds (94–103)* by Hanna Gnatchuk tests some hypotheses resulting from the synergetic view. For the relationship between word length and number of its compounds using *Longman Exams Dictionary* she obtained the Zipf-Alekseev function; for the relation between age of word and the number of its compounds she obtained the exponential function; for the relation between polysemy and number of compounds of a word she obtained the Lorenz function; and for the tendency of a word to form more compounds if it belongs to more than one POS she obtained the simple power function. The testing has been evaluated by means of the determination coefficient which was in all cases very high. The study is, as a matter of fact, a part of synergetic investigation specified for compounds.

Another synergetic article is *Context-Specific Distribution of Word Meanings (104–112)* by Maria Rukk studying the hypothesis: The higher the polytextuality is, the stronger the tendency is for the word to be used context-specifically. The author has chosen 15 Russian fiction texts and the polytextuality of the prepositions. In many languages they do not have to be lemmatized. The same can be done e.g. with the Ugro-Finnic suffixes. The results are presented in a table which can be used also for further research. Of course, further text should be

analyzed. Perhaps it would be interesting to compare the Slavic languages based on the same (translated) texts. The article shows especially the Russian research devoted to this problem.

While in texts we have to do only with one dimension, dialectology encompasses at least two. The enormous and numerous problems associated with this discipline are discussed by Hans Goebel and Pavel Smečka in the article *The Quantitative Nature of Working Maps (WM) and Taxatorial Areas (TA). A Brief Look at two Basic Units of Salzburg Dialectometry (S-DM)* (113–127). The domain is full of boundary conditions – frequently connected with historical phenomena – which can be taken into account with great difficulties. Thus the number of laws derived in this domain is, preliminarily, not abundant. The authors describe the whole discipline of dialectometry, bring several figures for which functions could be derived. The discipline was born in nineteenth century, developed in several European states and seems to be developing slowly: if one considers the labor necessary to set up a dialectological atlas – as compared with the analysis of one sole text – one must be astonished at the fact that linguists devoted their time to create them.

The problems concerning the same discipline are further analyzed by Sheila Embleton, Dorin Uritescu and Eric S. Wheeler in the article *Play with the Data!* (128–134) demonstrating the elasticity of dialectological data, proposing some methods (distance matrices, correlation, etc.), showing examples from Finnish and Romanian and finishing the article by a saying which should be learned by heart by all linguists: *Play with the data! One never knows what one might discover.*

The New Year messages of presidents are a preferred theme of text analysts. The text, often written by professional writers, shows usually the evolution of the state of affairs, tasks of economy and politics, i.e. the situation of the nation. This is a common subject of all speeches, hence a preferred text type. The Italian (and many other) presidential speeches have been analyzed earlier, the authors of the next article, Michele A. Cortelazzo and Arjuna Tuzzi analyze in *The First End-Of-Year Address by the New President of the Italian Republic Sergio Mattarella* (135–149) the lexical profiles of all available messages using Labbé's intertextual distance, cluster analysis and correspondence analysis. What is the place of the new president Mattarella? What are his differences to other presidents? A number of computations and figures are presented here, enabling other researchers either to evaluate the Italian situation or look at the messages of their own presidents.

Another synergetic study is the article *Thematic Concentration and Vocabulary Richness (150–159)* by Miroslav Kubát and Radek Čech. Their aim is to show that these two properties are negatively correlated. They are defined in

such a way that thematic concentration automatically causes small vocabulary richness. For measurement they use the secondary thematic concentration, the moving average – which seems to be a good instrument against too strong oscillation – the type-token ratio, McIntosh's the relative repeat rate, analyze 516 English texts in six text types and give a lot of references concerning this object of investigation. The Czech authors developed also a software QUITA¹ which computes all the necessary indicators and R. Čech analyzed Czech texts separately. This view of texts is well documented in the literature.

The approach proposed by Solomija Buk and Andrij Rovenchak in the article *Probing the "Temperature" Approach on Ukrainian Texts: Long-prose Fiction by Ivan Franko (160–175)* seduce the linguists to change the technique used in quantitative linguistics and try to interpret a function (Bose-function) used in physics (the second author is physicist) in linguistic terms – if one finds data following the function. They analyze the frequency spectrum of words, fit to it the Bose distribution in which one of the parameters says something about "temperature" in physical domain, and search for its links to other textual properties. It is a pity that they do not show the mathematical-linguistic background of the distribution though it has already been used several times in linguistics. In the article one finds the epithetization index (number of adjectives divided by the number of nouns in a given text), text size, vocabulary richness, year of publication, all applied to the Ukrainian author Franko. They touch the full texts, the direct speech, author's speech, language analytism etc. Evidently, they move in the centre of text analysis but one cannot tell everything in a short article. Besides, it would be, perhaps, better to adhere to a secure linguistic approach: (a) define, (b) set up a linguistic hypothesis, (c) translate it in mathematics and solve, (d) collect data, (e) test the hypothesis having the form of an indicator, a function or a distribution, (f) translate the result of the test into linguistics. In the meantime, ascribe to each parameter a linguistic property or a linguistic requirement (speaker/hearer) and you are in the centre of synergetic linguistics.

Since the "competent speaker" and the "competence" died several decades ago, one analyzes rather texts. The next article *Can Pronouns Change the Dynamic Visualization of the Poetic World? (176–182)* by Sergey Andreev is concerned not only with the role of pronouns in texts by Pushkin but shows also another way of constructing Köhlerian motifs. The study presents implicitly the enormous field of texts analysis and stimulates to quite new vistas. As a matter of fact, even the shortest text has an infinite number of properties. Andreev shows explicitly some of the ignored ones in poetry.

¹ Available at: <http://oltk.upol.cz/software>

Fan Fengxiang studies in his article *A Study on Segmental TTR, Word Length and Sentence Length* (183–195) three well known properties in 285 English journalistic texts but uses the techniques proposed by R. Frumkina, namely the subdivision of the text in passages. Fan divides the texts in three equal passages enabling him to study also the development of the given properties in the course of the text. He analyzes also the relationships between these properties and comes to the function $y = a \cdot x^b \cdot \exp(c \cdot x^d)$ which cannot be easily interpreted linguistically. Though the fittings are excellent, it has also the disadvantage that its differential equation deviates linguistically from the general theory and one does not know how to distribute the language-, writer- and reader-requirements. A further disadvantage is the fact that the parameter c acquires a rather microscopic size. Nevertheless, the study of texts in form of Frumkina passages seems to be a good variant to the use of moving averages.

Since poetic language has a very regular form, its study is a favorite research object of text analysts. Perhaps the last overall study was “Quantitative Analysis of Poetic Texts” by Popescu, Lupea, Tatar and Altmann (Berlin/Boston: de Gruyter: 2015). The number of known aspects is enormous and can be studied separately. In the next article *Statistical Analysis of the Diachronic Development of Terminal Rhyme in Chinese Poetry* (196–216) the authors, Xiaping Pan and Haitao Liu, scrutinize the Chinese rhyme in various types of Chinese poetry (ancient poetry written in the Tang and Song dynasties, modern Chinese Mandarin poetry written since the 1920s, modern Taiwanese poetry and Chinese translations of English poems). They state that the phonic form of the rhyme develops. This is caused by the fact that the repeating of old forms does not evoke new associations. Hence one can set up periods of poetry which may be different for every property. On the other hand, special features of rhyme may develop differently. The spectrum of this study, if extended onto many languages, may lead to a world-wide project.

The next text-analytic contribution *Lexical Text Compactness with Link Length Taken into Account* (217–227) written by two mathematicians, Gejza Wimmer and Ján Mačutek, who are intensively engaged in linguistics, chooses a different view, something between Belza-chains and hreb-analysis. They consider also the distance between two occurrences of the same lexeme – in terms of the number of intermediate sentences – and several other aspects as properties which can be used to express the lexical text compactness. The study is fully mathematized, as can be expected with mathematicians, and shows a new view of this very fuzzy problem. They analyze 59 Czech texts by the writer Karel Čapek written in 7 text types, evaluate and compare the results. What’s more, one can obtain the computation program from one of the authors.

It is very valuable and pleasant to see quantitative analyses of languages which are not the “daily bread” of normal linguists. One obtains the

impression that the linguistic reality is unique. In the article *Continuous Modelling of Verse Lengths in Welsh and Gaelic Metrical Psalmody* (228–236) Andrew Wilson shows that verse length in Welsh and Scottish Gaelic behaves in the same way as in all other modern cases, namely it follows the Zipf-Alekseev function. The author emphasizes that discrete or continuous modeling are merely our approaches – not the reality – and state that a simpler function, e.g. the power function, would, perhaps, be more adequate because the parameter c in the Zipf-Alekseev function takes on negligible small values. In all cases one should adhere to the principle that the applied functions should have as few parameters as possible. The elimination of some parameters from the general formula is caused by boundary conditions which need not be preliminarily explicated. The Zipf-Alekseev formula is a general scheme having several special cases.

Kamil Stachowski, in the article *German Loanwords in Polish and Remarks on the Piotrowski-Altmann Law* (237–259) shows a number of phenomena concerning the loans of German words in Polish and presents the numbers of loans ascribing them the century of adoption. In the given form they cannot follow the Piotrowski-Altmann law which concerns merely cumulative data. Nevertheless, the study is a good basis for testing the form of the law if the individual numbers would be given in numerical, not in graphical form.

The same problem is touched in the article *Probleme der Modellierung von Lehnbeziehungen (am Beispiel von Serbokroatismen im Slowenischen)* (260–272) by Emmerich Kelih and Ján Mačutek, concerning two Slavic languages. The authors have several problems with the law, look at it from various points of view and discuss it. Evidently, the discussion is not finished and one must take many boundary conditions into consideration. There is a difference between change and loan, one will, perhaps, be forced to consider the boundary conditions already in the differential equation.

Both last articles are excellent contributions stimulating to further study of a phenomenon which is law-like – as everything – but has a number of ad hoc circumstances which must be taken into account.

The last article of the volume, *Fast Calculation of Entropy with Zhang's Estimator* (273–285) written by a team of Spanish authors, Antoni Lozano, Bernardino Casas, Chris Bentz, Ramon Ferrer-i-Cancho is rather mathematical in its form. For concrete computations the authors use the Parallel Bible Corpus and the Universal Declaration of Human Rights but this is used only to show the way of computation, not the character of individual languages. It is to be hoped that in the future the authors will take articles from one language, of one text type and appeared in the same year in order to show the importance of entropy as described in the first part of the article.

The volume presents the center of quantitative linguistics which gains more and more importance in the linguistic world in which mostly grammatical rules were respected. There are a number of problems which are partially solved and the next generation will search for other ones as is usual in science. However, there is no better forum to discuss quantitative linguistic problem than in the volume dedicated to Reinhard Köhler.