Book review

Language and Text. Data, Models, Information and Applications. By Pawłowski, A., Mačutek, J., Embleton, S., Mikros, G. (Eds.). Amsterdam/Philadelphia: John Benjamins Publishing Company. 2021.

Emmerich Kelih [1*]

[1] Institute for Slavonic Studies, University of Vienna
[*] Corresponding author's email: emmerich.kelih@univie.ac.at

DOI: https://doi.org/10.53482/2022_53_403

In 2021 the omnibus volume *Language and Text. Data , models, information and applications* appeared in the prestigious *Current Issues in Linguistic Theory* series as volume 356. The volume contains a selection of papers presented at the 10th Qualico in Wrocław, Poland, in 2018. It has 17 individual contributions, which are presented to the reader in two subsections (*I. Theory and Models*, II. *Empirical Studies*), complemented by a brief introduction and a subject index at the end. Moreover, at the end of the volume, a short memorial contribution to the eminence of quantitative linguistics researcher Gabriel Altmann (1931–2020) is given by some members of IQLA (International Quantitative Linguistics Association, http://iqla.org/).

The introduction (given by the editors of the volume) provides a conceptual embedding of the published papers into some recent trends of quantitative linguistics (QL) and digital humanities. The current state of the art, as reflected by the papers, is on the one hand characterized by "traditional" QL research topics like well-known quantitative laws (for instance Zipf's, Menzerath's, and Piotrowski's laws), stylometry, and dialectometric studies. But as highlighted by the editors, on the other hand, recent ongoing changes in text production and text analysis (in particular the impact of digitalization, the availability of massive text corpora and the possibility of automatic data extraction, "big data" etc.) make the application and discussion of new methods of quantitative and automatic analysis quite necessary and reasonable. QL doesn't ignore these recent trends and this omnibus volume, in particular the second part, reflects these ongoing changes in a fast-changing research field.

The first contribution, "On the impact of the initial phrase length on the position of enclitics in Old Czech", a joint publication by Radek Čech, Pavel Kosek, Olga Navrátilová, and Ján Mačutek, seemingly reflects the traditional realm of QL. Quite the contrary is true, since a traditionally less researched

subject, namely historical texts from Old Czech (translations of the Bible), is in the focus of the analysis. Linguistically the word order and in particular the positioning of enclitics within the clause are examined. Based on correlation analysis some overall trends could be confirmed, where with ongoing length of the clause a decrease of the proportion of enclitics in post-initial position is observed. Although the results are based on three selected enclitics only, the given preliminary results ask for a future cognitive interpretation of such fine-grained positioning of enclitics. The paper by Lars G. Johnsen, "Term distance, frequency and collocations", deals mainly with a collocational analysis of about 440,000 books from the Norwegian National library. Two methods, based on the frequency of target words and the distance between the target word and its collocation, are presented, although as it is pointed out by the author, syntactic constraints also have to be considered, since they determine how close one word can come to another. The paper "A method for the comparison of general sequences via type-token ratio" by Vladimír Matlach, Diego Gabriel Krivochen, and Jiří Milička is also devoted to new methodological approaches. Here the analysis and comparison of strings on a very general level is discussed, although they are demonstrated mostly by means of linguistic data (*n-grams*). More traditional fields of analysis are again covered in the next paper, "Quantitative analysis of syllable properties in Croatian, Serbian, Russian and Ukrainian", by Biljana Rujević et al. For these four Slavic languages the syllable frequencies (the syllabification is performed automatically) and syllable length is calculated and the related raw data are modelled by appropriate statistical models. Moreover, the authors present a generalized version of Altmann–Menzerath's law for modelling the word length – mean syllable length relation also on the tokens level. The analysis is carried out based on the parallel text corpus used, which is in an important complement to "traditional" dictionary-based syllable frequency analyses. In her paper "N-grams of grammatical functions and their significant order in the Japanese clause" Haruko Sanada investigates n-gram frequencies of grammatical function types in Japanese noun phrases and the position of adjuncts expressing time, place or occasion. By using appropriate statistical tests some preferences towards an indeed free word order setting for Japanese can be shown. Petra Steiner's contribution "Linking the dependents. Quantitative-linguistic hypotheses on valency" relates syntactic and semantic aspects of case and valency to morphological properties, i.e. a new synergetic circle (in fact already well-reflected hypotheses are offered) of mutual interrelations in the sense of Köhlerian synergetic linguistics is offered. Relja Vulanović comes up in his paper "Grammar efficiency and the One-Meaning-One-Form principle" with particular problems of grammar efficiency based on various part-of-speech systems. Different types of grammar-efficiency formulas developed by the author in the past are compared and new ones are offered. The final paper of Part I is the contribution "Distribution and characteristics of commonly used words across different texts in Japanese" by Makoto Yamazaki. In this paper the interesting question is tackled to what extent high-frequency lexemes of Japanese assume an expected "Zipfian behaviour" or not. Moreover, it is discussed which role the text length and the occurrence in different texts (diversification, in synergetic linguistics also named *polytexty*) hereby play. The related empirical results are based on a large balanced corpus of contemporary written Japanese.

The second part, "Empirical Studies", starts with the brief contribution "The perils of big data" by Sheila Embleton, Dorin Uritescu, and Eric S. Wheeler. The advantages, disadvantages, and arising problems when working with big data (becoming available due to the digital turn in linguistics) in dialectology, a field where the authors have gained rich experiences in the last decades, are discussed. The paper "From distinguishability to informativity. A quantitative text model for detecting random texts" by Maxim Konca et al. can be positioned on the intersection of computational and quantitative linguistics and brings interesting insights into the structure of random and non-random texts through the lens of a bulk of quantitative text characteristics as discussed in recent quantitative lexicology (among others, h-point, various interpretations of the repeat rate, entropy, TTR measures, etc. are used). George Mikros and Rania Voskaki report in their contribution "A Modern Greek readability tool. Development of evaluation methods" about recently developed readability analysis tools which aim to help improve the judgements about the readability of Greek texts. The tools are based on various stylometric indices which were recently developed in quantitative text analysis. Methodological machine-learning and random-forest methods were used as proper tools. A novel research field for quantitative linguistic studies is introduced by Jiří Milička and Alžběta Houzar Růžičková in their contribution "Phonological properties as predictors of text success". Their idea is to explore to what extent reader-based "aesthetic" judgements of written texts interrelate with selected phonetic/phonological features responsible in Czech for euphony effects. A contribution to the recent stylometry of political speeches is given by Michal Místecký in his paper "Calculating the victory chances. A stylometric insight into the 2018 Czech presidential election". The paper is written in the tradition of quantitative presidential speech analyses and gives deeper insights into the style and language of Czech presidency candidates. Hermann Moisl's paper "Topological mapping for visualisation of high-dimensional historical linguistic data" offers new and innovative methods for the visualization and clustering of high-dimensional data, whereby he focuses on a mostly neglected, but rather important issue of the non-linearity of linguistic data. In the papers "Book genre and author's gender recognition based on titles. The example of the bibliographic corpus of microtexts" (written by Adam Pawłowski, Elżbieta Herden, and Tomasz Walkowiak) and "Quantitative analysis of bibliographic corpora: Statistical features, semantic profiles, word spectra" (the authors are A. Pawłowski, Krzysztof Topolski, and E. Herden) bibliometric data analysis is combined with contemporary QL, computational linguistics, and artificial intelligence approaches. Selected hypotheses regarding the structure of book titles (useful for automatic text genre detection) and the gender of authors are presented and tested on big data, coming from huge Polish bibliographic resources. In addition, in the second paper, the frequency of parts of speech and word spectra are studied and modelled by appropriate statistical methods. The last paper of Part II is the paper "Analysis of English text genre classification based on dependency types", written by Yaqin Wang. Thematically the paper is devoted to selected classification problems, based on a syntactic input (frequency of dependency types in the British National Corpus). Three different classification methods (PCA, hierarchical clustering, and random forest) are used and the related results are compared.

Overall, the omnibus volume can be recommended to every reader interested in the recent research output in QL and closely related fields. Although the volume appeared three years after the Qualico 2018 conference, there is no loss in relevance and topicality of the presented research. Quite the contrary seems to be true – quality requires time, and the volume represents a good example of this in respect to the theoretical and methodological level of most of the presented papers and also in regard to the style, the layout, and language of the contributions. Most of the papers are co-authored, which is of course quite natural because of the intrinsic interdisciplinarity of QL, where in almost all cases a linguistic background, computational skills, and statistical competence is required.