Review of Quantitative Approaches to the Russian Language

Article · May 2018

CITATIONS
0

READS
64

1 author:

Heng Chen
Guangdong University of Foreign Studies
14 PUBLICATIONS **32** CITATIONS

SEE PROFILE

Glottometrics 41, 2018, 91-95

Mikhail Kopotev, Olga Lyashevskaya, & Arto Mustajoki. (Eds.) (2017). *Quantitative Approaches to the Russian Language*. New York: Routledge. ISBN:978-1-138-09715-5, 220 pp.

Reviewed by Heng Chen[1]

Russian linguistics has a tradition of quantitative/mathematical linguistic studies, which can be dated back to the 19th century. Quantitative Linguistics (QL) in Russian had been developed synchronously with international QL studies. There was actually a boom for Russian quantitative studies in the 1960s-1980s, the most famous of which, including Piotrovsky's "Statistika Reći" ("Parole Statistics") group, Tuldava's series of quantitative studies regarding lexical systems, and Arapov's work of *Quantitative Linguistics,* etc. The experts in the "Parole statistics" group are not only from linguistics, but also other disciplines such as computer science, mathematics, psychology, and statistics, etc. However, the QL studies merely vanished after the end of the Soviet Union, although there are still several excellent QL researchers such as B.B. Kromer and A.A. Polikarpov. Kelih (2008) conducted a more systematic historical investigation of the application of quantitative methods in Russian linguistics and literature science, for a review, see Liu (2010).

Nevertheless, many large and deeply annotated corpora are available for extensive quantitative studies nowadays, such as the Russian National Corpus, ruWac, and ruTenTen, just tp name a few. Most of these articles in this volume are achievements of a workshop entitled *Quantitative Approaches to the Russian Language*, which took place in August of 2015 in Helsinki, Finland, co-organized with a symposium called New Developments in the Quantitative Study of Languages. This volume is a new attempt in this field by applying the latest new techniques such as NLP tools, mathematical models, and machine learning algorithms to quantitative analyses of Russian big language data, meanwhile, the methods are also evaluated. This volume is focused on quantitative methodology and data processing of Russian language, representing state-of-art research in Russian QL. There are ten articles in this volume including the first introduction chapter, which are organized into four parts around the following topics:

Part I, Introductory chapters, including 2 contributions, opens with an introductory article titled "**Russian challenges for quantitative research**" by **Mikhail Kopotev**, **Olga Lyashevskaya**, and **Arto Mustajoki**, who are also editors of this volume. The authors begin by stating that the goal of the present volume is to present current trends in examining Russian QL, to evaluate the new research methods and modals vis-a-vis Russian data, and to show the advantages and disadvantages of the methods and models. Then they describe the

[1] Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China. Email: chenheng@gdufs.edu.cn

main features of the Russian language and look back upon the quantitative (corpus) studies in Russian (2000-2010s), concluding that many topics in grammar and lexicon need to be covered, and more examples of quantitative approaches need to be provided. Next, the contributions in this volume are introduced. The inventory of internet sources and quantitative methods used in this volume are summarized at the end, which makes it advantageous for the inquiry.

The other contribution in this part, "**Big data and word frequency: measuring the consistency of Russian corpora**" by **Maria Khokhlova,** aims to compare linguistic phenomena across the main Russian corpora of different sizes. Specifically, 3 linguistic phenomena are examined, i.e., syntactic relations involving nouns, high-frequency nouns, and low-frequency nouns; the corpora include the ruWac and the ruTenTen; the main quantitative methods are log-likelihood score and Spearman's correlation coefficient. The results obtained for the syntactic relations involving nouns in ruWac and ruTenTen are compared with each other, and the analyses show that the two corpora are largely similar in featuring syntactic relations. The results of the high- and low-frequency Russian nouns were compared with data published in *A Frequency Dictionary of Modern Russian*, which indicate that there are different situations for high- and low- frequency distributions comparisons. Further researches are needed for more parts of speech and other better metrics.

Part II, Topics in semantics, to be more specific, lexical semantics, contains 3 contributions. It begins with an article titled **"Looking for contextual cues to differentiating modal meanings: a corpus-based study"** by Olga Lyashevskaya, Maria Ovsjannikova, Nina Szymor, and Dagmar Divjak. An important property of modal words is that they are largely ambiguous. Thus the modals can be assumed to be "word-like elements which are poly-functional in the sense that they express no less than two types of modality". The authors propose that the availability of large corpus data paves the way for a study of the empirical reliability of existing classifications originally proposed by philosophers. Thereupon, in order to test if contextual cues, i.e., 12 formal and semantic features (of the modals) can predict the type and function of modal words, the most frequent 6 Russian verbs were chosen, and for each word, 250 sentences were extracted from the RNC. The annotation of contextual cues for each word in the sentences was done by two experts manually. To achieve the aim, two visualization techniques, i.e., multiple correspondence analysis and shaded mosaic plots, and two inferential statistical methods, i.e., polytomous logistic regression, and classification and random forest were used. The results show that, generally, type or function can be predicted from context cues, also with some exceptions, which need further investigations in the future.

The study titled **"Automated word sense frequency estimation for Russian nouns"** by Anastasiya Lopukhina, Konstantin Lopukhin, and Grigory Nosyrev, is the first study on sense frequency distributions in the Russian language. The article begins with a well-known observation by G. k. Zipf (1945), stating that words used more frequently usually have more senses than words that are used less frequently. Although information about word frequency is widely available nowadays, sense frequencies and their distributions remain a neglected area in linguistics. In this paper, the authors present a method for determining noun sense frequency distributions automatically from raw text, an evaluation of the methods, its comparison to state-of-art system, and a discussion of its applications. The method is actually

based on word sense disambiguation techniques usually used in computational linguistic or NLP, using distributed vector representations with weighting. Distributed vector representations is a way of representing words as low-dimensional dense real-valued vectors, and is known as the famous word2vec family of methods. The linguistic hypothesis here is that words occur in similar contexts tend to have similar meaning. The evaluation results show that the frequency estimation error of the model is 11-15 percent. The results of the 440 nouns sense frequency information as well as source code are online for further consultation.

The third contribution in this part is **"Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes" by** Andrey Kutuzov and Elizaveta Kuzmenko. Similar to the above research by Lopukhina et al, this study traces Russian word semantic changes with state-of-art technique in lexical semantic similarity modeling: artificial neural networks (neural embedding models) in NLP. The central assumption here is that online training of such models with new textual data results in a "drift" of word vectors in the "semantic space". The case presented in this study uses three sub-corpora from the RNC: texts produced before Soviet times (before 1917), during Soviet times (1918-1990), and after the fall of the USSR (since 1991). After training 3 neural embedding models on these 3 subcorpora, several algorithms to extract words with changing meanings are evaluated. Eventually, they came to conclusion that comparing nearest neighbor sets using Kendall's τ distance works best, both on artificially created data and on short, manually compiled, gold standard data sets. The results of 2000 nouns and adjectives that have undergone the most significant changes are online for further consolation.

PART III, Topics in the Lexicon-Grammar Interface, including 3 contributions, begins with **"The grammatical profiles of Russian biaspectual verbs"** by Alexander Piperski. Biaspectual words can be used to convey both perfective and imperfective meaning. In this study, three quantitative methods for determining the status (more imperfective or perfective-like) of biaspectual verbs (over time) were evaluated: estimating the relative frequency of their perfective and imperfective gerunds, classifying their grammatical profile using the *k* Nearest Neighbors algorithm, and conducting an experiment on the perception of the inherent aspect of biaspectual verb forms. The results show that their applications are in agreement with each other.

A study conducted by Lidia Pivovarova, Daria Kormacheva, and Mikhail Kopotev, titled **"Evaluation of collocation extraction methods for the Russian language",** begins with a distinction between lexical and empirical collocations, of which the later is the focus of this study. Then the authors review the main existing measures for collocation extraction, including t-score, log-likelihood, mutual information, Dice, and wFR. Next, the evaluation of automatically obtained collocations is conducted by comparing both with dictionary data and native speakers' responses. Two comparisons both show that t-score performs slightly better than the other measures. However, they all provide similar results, which means that it may be more plausible to suppose that different measures are intended to identify different kinds of collocates.

The third contribution in this part is **"From quantitative to semantic analysis: Russian constructions with dative subjects in diachrony"** by Anastasia Bonch-Osmolovskaya. The author conducts a quantitative research into predicative and corresponding adjective constructions with dative arguments from a diachronic perspective. The core issue here is to

reveal behavior classes of lemmas defined in terms of dative argument frequency within the three forms (i.e., predicative, short adjective form, and long adjective form) and to study diachronic changes of the determined behavior classes. The data are from the RNC and the search is confined to two samples, one from the 18th century, the other from the 21th century. Eight lemmas are selected for the study. The investigation shows that the frequency rates of dative subjects are different from predicates, and diachronic trends are observed using hierarchical clustering methods.

PART IV, also the final part, turns our attention to **Topics in language acquisition,** including 2 contributions**. "Measuring bilingual literacy: challenges of writing in two languages"** by Aleksei Korneev and Ekaterina Protassova. This study focuses on a computer-based, contrastive assessment of bilingual Finnish-Russian primary students with different linguistic backgrounds, and examines their written language proficiency. To achieve this, experiments of fours groups - the Russian Dominant Bilinguals with 15 children, the Finnish Dominant Bilinguals with 13 children, the Russian-speaking control group with 15 children, and the Finnish-speaking control group with 10 children - are conducted based on a computer handwriting assessment system. The handwriting parameters include mean time of writing a letter, the exact time to write separate letters, the stability of the edge of the line. To analyze the parameter differences among different groups of subjects and in different writing tasks (copying and dictation), the authors use the repeated measures ANOVA. The results show that the dominance of the language plays an important role in writing proficiency in bilinguals; the writing system is another important factor; the language of the environment might support the language skills, but training in a different language and in a different script supports the quality of writing.

The final contribution, **"When performance masquerades as comprehension: grammaticality judgments in experiments with non-native speakers" by** Robyn Orfitelli and Maria Polinsky. In language acquisition studies, many observations are based on experiments. However, inappropriate experimental design can be problematic, because it can be hardly replicated and re-examined. In this study, the authors criticize the grammaticality judgment tasks (GJTs) which are originally introduced in linguistics to measure the acceptability of particular language structures for native speakers, and are now misused for non-native speakers. Based on numerous instances of within- and across-task inconsistency, the authors argue that the metalinguistic demands imposed by the task - and the difficulty involved in identifying the root cause of any incorrect answers - render the task unsuitable for testing language comprehension with non-native speakers. Then the authors illustrate this problem by discussing two recent experiments conducted with Russian non-native speakers using GJTs and other tasks. The analysis suggests that poor performance on GJTs by non-native speakers may be related not to grammatical errors, but to extra-grammatical factors involving metalinguistic awareness and processing demands.

In conclusion, this edited collection presents a range of resources and new quantitative methods in Russian QL studies, which will promote the combination of classical QL with the latest techniques from the age of language big data. The authors show that those state-of-art techniques such as neural embedding models, word2vec, word sense disambiguation (WSD) algorithms and distributional semantic models, actually can and should be applied to quantitative studies of Russian language regarding modern linguistic questions. Moreover, a

series of evaluations of quantitative methods are conducted, and some theoretical problems are also scrutinized.

This volume was published with high typographical quality, and the index listed at the end of the book makes it very convenient for readers to read and refer. However, there are still a few critical points I am obliged to make. The "R2" in pp. 60, pp. 66, and pp. 72, should be "R^2" or with the superscript "2"; the "Diagram 1" in pp. 168 should be "Figure 8.3", and the "Figure 8.3" in pp. 169 should be "Figure 8.4", otherwise it will be confusing. As for this collection, we believe that it would be greatly improved if it focuses more on quantitative linguistic laws (Köhler, 2012), and gives more in-depth linguistic interpretations and predictions. What is more, we are looking forward to more contributions in topics such as quantitative syntax analysis, linguistic complex systems/networks analysis, which are research focuses in QL today.

The Russians have contributed a lot to the development of QL as well as computational linguistics, for example, the most famous Markov Chain extensively used and developed in NLP practices, the Piotrowski-Altmann law in QL (one of the three main laws in QL), as well as the precise literary studies by B. I. Yarho that can be dated back to the early 20th century. Now, by applying the state-of-art techniques to Russian QL studies, this volume will promote developments in all these fields. I think this will be of great interest to graduate students and researchers in the area of quantitative and Slavic linguistics, both outside and inside Russia.

References

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin and Boston: De Gruyter Mouton.

Liu, H. (2010). Review of Kelih, Emmerich (2008) *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft* (*History of the application of quantitative methods in Russian linguistics and literature*). Hamburg: Kovač. *Journal of Quantitative Linguistics,* 17(4): 365-370.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2), 251–256.