

BOOK REVIEWS

Quantitative Approaches to the Russian Language, edited by Mikhail Kopotev, Olga Lyashevskaya, and Arto Mustajoki, London, Routledge, ISBN 9781138097155

The omnibus volume under review comprises 10 individual chapters by 22 authors, thus most of the chapters are co-authored. This seems to reflect the overall interdisciplinary approach focus of the volume, most of the articles in which originate from a workshop entitled ‘Quantitative Approaches to the Russian Language’, held in Helsinki, Finland in August 2015.

Part I (‘Introductory Chapters’) starts with the contribution ‘Russian Challenges for Quantitative Research’ by Mikhail Kopotev, Olga Lyashevskaya and Arto Mustajoki. It begins with a short linguistic overview of the main features of the Russian language, followed by an extensive bibliographical sketch of quantitative (however mainly statistical corpus) studies in Russian (from 2000 through to the 2010s). Alongside the available corpus sources, a range of statistical methods that are currently used and discussed in corpus linguistics are presented (significance testing, multivariate methods, various regression models, etc.). In the final section all the published contributions are briefly introduced, again from the perspective of the used data sources and applied quantitative methods. The second contribution, ‘Big Data and Word Frequency: Measuring the Consistency of Russian Corpora’ by Maria Khokhlova, discusses in general terms the (old and often discussed) problem of representativity, based on the size of corpora analysed. In comparing three different Russian corpora the author aims to identify differences and similarities in the frequency of lexical items, in particular for high- (and low-) frequency nouns and so-called syntactic relations (bigram analysis on the level of parts of speech and morphological markers, e.g. adjective + noun, verb + dative constructions, etc.). The statistical comparison is based on Spearman’s rank coefficients, and it can be shown, for example, that high-frequency nouns coincide much more than low-frequency items. Although a theoretical perspective is absent from the paper (for example, why would differences and similarities be expected in this kind of linguistic ‘big data’ at all?), the empirical results per se are interesting.

Part II (‘Topics in Semantics’) starts with the contribution ‘Looking for Contextual Cues to Differentiating Modal Meanings: A Corpus-Based Study’, written by Olga Lyashevskaya, Maria Ovsjannikova, Nina Szymor and Dagmar Divjak. The paper is devoted to Slavic (in particular Russian) modal words and their classification along certain semantic types (e.g. deontic, participant-external, participant-internal, epistemic), their function, their semantic class of the predicate under modality, and so on. The attractiveness of this contribution lies in the application of various quantitative methods (including multiple correspondence analysis) to the analysis and the visualization of modality data coming from relevant corpus sources of Russian. Particular

focus is placed on determining the contexts in which a modal word would prompt a listener to interpret a given item as modal. While the basic idea of the paper seems to be clear, the three presented ‘hypotheses’ (p. 52) about the evocation of modal words by the logical type of modality and/or the interpretation of the modal function and the relevance of the context (in particular Hypothesis 3, on the dependency of modality interpretation on external knowledge about the world and the situation) are mostly assumptions and in some way constitute the naming of influence factors, but they are by no means formulated as empirically testable hypotheses – and, in fact, no testing of these given ‘hypotheses’ follows. What does follow however is a (dense) application of multiple visualization techniques, polytomous regression models, analyses of variance (ANOVAs) and random forest methods. In the final part no clear-cut results are yielded – at least none referring to the hypotheses presented in the introduction.

The next paper in Part II, ‘Automated Word Sense Frequency Estimation for Russian Nouns’ by Anastasiya Lopukhina, Konstantin Lopukhin and Grigory Nosyrev, is devoted to questions of word senses of Russian nouns, i.e. the polysemy of linguistic units. The authors propose a model based on semantic vectors, trained on word contexts from large corpora. The main approach is based on the well-known idea that words occurring in a similar context tend to have a similar meaning. Noteworthy is the inclusion of the recently developed *Active Dictionary of Russian*, where lexemes are described according to their semantic and syntactic properties, collocational restrictions and synonyms and antonyms. The authors provide the full data from their research (<http://sensefreq.ruslang.ru/>), and the relative frequency of different word senses for Russian is given. This data could serve as a starting point for modelling procedures as usually done in quantitative linguistics. This contribution, although rooted in computational linguistics approaches to word sense disambiguation, is a welcome stimulus for ongoing research in this area.

In one further paper, ‘Two Centuries in Two Thousand Words: Neural Embedding Models in Detecting Diachronic Lexical Changes’, Andrey Kutuzov and Elizaveta Kuzmenko deal with the semantic shifts of nouns and adjectives in Russian before, during and after Soviet times. Based on a distributional semantic approach (where word meaning is determined by context), various methods for calculating the overlap of the context of the studied lexemes are proposed. The Jaccard Similarity Index, Kendall’s τ and others are tested on a subset of over 40 Russian nouns for which, based on ‘qualitative’ linguistic analyses, the kind of semantic shift is already known. Bearing in mind the ongoing progress of building historical corpora, such approaches are highly welcome, and frequency data about semantic developments can provide new stimuli for quantitative linguistics about law-like language changes, as formulated in the well-known Altmann–Piotrovskij law.

In Part III (‘Topics in the Lexicon–Grammar Interface’), Alexander Piperski in his paper ‘The Grammatical Profiles of Russian Biaspectual Verbs’ suggests different quantitative methods (based on the frequency of perfective and imperfective gerunds or the frequency of tense or aspect mood forms) for the determination of the status of biaspectual verbs in Russian. In addition to the mass of

verbs, which are either perfective or imperfective, in Russian there is also a small number of biaspectual verbs; the author presents some methods which help in identifying them as 'more' perfective or 'more' imperfective. As a result, a 'likeness' of being perfective or imperfective can be detected, giving an interesting prognosis tool for future development that can be used in the analysis of other Slavic languages, like Croatian/Serbian and Slovene, which are also known for their large numbers of biaspectual verbs.

The next paper, 'Evaluation of Collocation Extraction Measures for the Russian Language' by Lidia Pivovarova, Daria Kormacheva and Mikhail Kopotev, begins with an exhaustive overview of the existing measures for collocation extraction in corpora. The overlaps of the various measures are discussed for selected cases, after which the lexical collocations under consideration are evaluated by native speakers, who are not always able to distinguish between idioms and lexical- and frequency-based collocations (p. 150). Moreover, many of the statistically gained collocations are not recognized as such by native speakers. Therefore, it appears that at the level of collocation the differentiation between 'objective' and 'subjective' probabilities clearly has a heuristic value, but at the same time automatically retrieved collocations do not have the status of 'fixed' expressions in the minds of the speakers.

In the next paper, 'From Quantitative to Semantic Analysis: Russian Constructions with Dative Subject in Diachrony' by Anastasia Bonch-Osmolovskaya, the use of dative subjects in predicative and adjective constructions is considered. The analysis is focused on shifts in the frequency of the mentioned constructions. Based on the Russian National Corpus, the main analysis tries to answer the question of whether the frequency of dative subjective constructions with overt dative arguments has changed over the last two centuries or not, depending on the type of construction. By means of hierarchical cluster methods and based on cluster size types, the author shows the extension and increased usage of such constructions in recent decades. However, it remains unclear as to whether or not this increase in the spectrum of the functionality could also be explained by the different amounts of data available for each period analysed.

In the final part ('Topics in Language Acquisition'), Aleksei Korneev and Ekaterina Protassova present in their paper 'Measuring Bilingual Literacy: Challenges of Writing in Two Languages' the results of a survey on the writing skills (in particular handwriting) of bilingual pupils (either Russian-Finnish or Finnish-Russian) in bilingual schools in Helsinki, also taking into consideration the reading proficiency in the children's homes. Lastly, in 'When Performance Masquerades as Comprehension: Grammaticality Judgments in Experiments with Non-Native Speakers', Robyn Orfitelli and Maria Polinsky deal with grammatical judgement tasks (GJTs) in order to explore the metalinguistic awareness of non-native speakers. They also present the results from a comparison with experiments on sentence-picture matching. Clearly, these last two contributions are not thematically located at the core of the omnibus volume and are either of 'local' interest or generally more methodologically orientated. Overall, the collection is coherently structured and well designed. However, in the references (pp. 135-136) one finds

a nonstandard upper- and lower-case writing in Russian (e.g. *v Sovremennom Russkom Jazyke vs Problemy I Rešenija*) and an unnecessary mixing of transliteration rules (s, sh), and in the Cataloguing in Publication (CIP) one finds an illegible writing of the Russian names of the editors, which is surely not the fault of the editors.

In sum, this volume gives a good overview of the current state of the art in empirical linguistics, relying heavily on the particular subset of statistical methods applied. However, the title of the volume is at least partly misleading, since any links to current studies on Russian and quantitative Russian linguistics are missing. The focus is clearly placed on (statistical) corpus linguistics, wherein – as this volume shows – statistical methods can doubtlessly be applied fruitfully, albeit mainly embedded as an inductive tool with which to obtain general empirical tendencies. Taking into account the rich tradition of Russian quantitative linguistics – represented by outstanding scholars such as R. G. Piotrovskij, M. V. Arapov, J. A. Tuldava and J. K. Krylov, among many others – this volume seems to represent a ‘new’ beginning of the application of quantitative methods of Russian, but by no means of post-Soviet and Russian linguistics in general.

Emmerich Kelih

University of Vienna

 emmerich.kelih@univie.ac.at

© 2019 Emmerich Kelih

<https://doi.org/10.1080/09296174.2018.1558834>