Computing Russian Morphological distribution patterns using RusAC Online Server

Galya Gatiyatullina
Department of theory and practice of language teaching,
Research laboratory 'Intellectual technologies for text management'
Kazan Federal University
Kazan, Russia
ggaliya-m@mail.ru

Marina Solnyshkina
Department of theory and practice of language teaching, Research laboratory 'Intellectual technologies for text management'
Kazan Federal University
Kazan, Russia
mesoln@yandex.ru

Valery Solovyev
Linguistic research and education center, Research laboratory 'Intellectual technologies for text management'
Kazan Federal University
Kazan, Russia
maki.solovyev@mail.ru

Andrey Danilov
Department of bilingual and digital education,
Kazan Federal University
Kazan, Russia
tukai@yandex.ru

Ekaterina Martynova
Research laboratory 'Intellectual technologies for text management'
Kazan Federal University
Kazan, Russia
katerinamarty@yandex.ru

Iskander Yarmakeev
Institute of Philology and Intercultural Communication,
Kazan Federal University
Kazan, Russia
ermakeev@mail.ru

Abstract—The article presents findings of distribution patterns of Russian grammatical categories computed with the help of MyStem.3 tagger and a proprietary Russian language processor, ETAP-3. The corpus of over 1.1 mln tokens compiled for the study comprises two types of academic textbooks used in Russian schools: Science and Humanities. We computed descriptive metrics of each textbooks with the help of the text analyzer RusAC (http://tykau.pythonanywhere.com/) and pursued the contrastive analysis of Science and Humanities textbook features. Significant differences of two types of the texts were found in distribution patterns of noun cases and verbs tenses, while morphological categories of nouns, adjectives, verbs, and adverbs demonstrate similarities. The specifics of grammatical patterns defined for classroom textbooks can be used in further studies on distribution of morphological patterns and text complexity of Russian academic texts.

Keywords—corpus linguistics, types of texts, genres, academic register, text complexity, morphological distribution, distribution patterns

REFERENCES

[1] J. H. Greenberg, "A quantitative approach to the morphological typology of language," International Journal of American Linguistics, vol. 26(3), 1960, pp. 178–194.

[2] H. Kucera and W. N. Francis, "Computational analysis of present-day American English," Providence: Brown Univ. Press, 1967, 424 p.

[3] D. Biber "Variation across Speech and Writing," Cambridge, UK: Cambridge University Press, 1988, pp. 299.

[4] T. Givon, "Markedness in Grammar: Distributional, Communicative and Cognitive Correlates of Syntactic Structure," Studies in Language. International Journal sponsored by the Foundation "Foundations of Language", vol. 15, Issue 2, Jan 1991, pp. 335–370.

[5] D. Biber, "University language: A corpus-based study of spoken and written registers," Amsterdam: John Benjamins, 2006, 261 p.

[6] D. Biber et al., "Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus", Report, ETS Research Memorandum, 2004, 374 p.

[7] D. Biber, "Register as a predictor of linguistic variation," Corpus Linguistics and Linguistic Theory vol. 8–1, 2012, pp. 9–37.

[8] E. Teich, S. Degaetano-Ortlieb, H. Kermes, and E. Lapshinova-Koltunski, "Scientific registers and disciplinary diversification: A comparable Corpus approach," in Proceedings of Sixth Workshop on Building and Using Comparable Corpora (BUCC), 2013, pp. 59–68.

[9] A. C. Fang, J. Cao, "Text Genres and Registers: The Computation of Linguistic Features," Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, 280 p.

[10] E. Teich, S. Degaetano-Ortlieb, P. Fankhauser, H. Kermes, and E. Lapshinova-Koltunski, "The Linguistic construal of disciplinarity: A data-mining approach using register features," Journal of the Association for Information Science and Technology, 67(7), 2016, pp. 1668–1678.

[11] D. Biber, S. Johansson, G.Leech, S. Conrad and E. Finegan. "Longman Grammar of Spoken and Written English," Harlow: Pearson Education Limited, 1999, 1204 p.

[12] D. Biber and B. Gray, "Grammatical complexity in academic English: Linguistic change in writing," Cambridge: Cambridge University Press, 2016, 276 p.

[13] A. C. Fang, "Verb forms and sub-categorisations," Literary and Linguistic Computing 12 (4), 1997, pp. 209–217.

[14] D. Biber and S. Conrad, "Register, Genre, and Style," 2nd edition, Cambridge: Cambridge University Press, 2019. 420 p.

[15] A. F. Zhuravlev, "An experience of quantitative and typological investigation of spoken registers" [Opyt kvantitativno-tipologicheskogo issledovaniya raznovidnostey ustnoy rechi], Varieties of urban spoken language: a collection of research articles – Raznovidnosti gorodskoy ustnoy rechi, Moscow, Nauka, 1988, pp. 84–150.

[16] O. B. Sirotinina, "Spoken language within the system of functional styles of the Russian literary language: grammar" [Razgovornaya rech v sisteme funktsionalnyh stiley sovremennogo russkogo literaturnogo yazyka: grammatika], 3rd edition, Moscow, Librekom, 2009, 312 p.

[17] M. Solnyshkina, V. Solovyev, V. Ivanov, A. Danilov "Studying Text Complexity in Russian Academic Corpus with Multi-Level Annotation", Computational Models in Language and Speech, Workshop (CMLS 2018) Proceedings , 2018, Vol.2303, pp. 93–103.

[18] D. Biber, "Representativeness in Corpus Design", Literary and Linguistic Computing, vol. 8 (4), 1993, pp. 243-57.

[19] A. K. Asiryan, "Morphological tagging tools comparison", Intellectual potential of the XXI century, 2017. URL: https://www.sworld.com.ua/konferu7-317/27.pdf.

[20] O. V. Dereza, D. A. Kayutenko, A. S. Fenogenova "Automatic morphological analysis for Russian: A comparative study", Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies. Student session (online publication), 2016. URL: http://www.dialog-21.ru/media/3473/dereza.pdf.

[21] L. Iomdin, V. Petrochenkov, V. Sizov, L. Tsinman, "ETAP parser: state of the art", 2012. URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Iomdin.pdf

[22] V. Solovyev, V. Ivanov, M. Solnyshkina, "Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics", Journal of Itellegent and Fuzzy Systems, Vol.34, Is.5, 2018, pp. 3049–3058.