Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

EFFICIENCY OF TEXT READABILITY FEATURES IN RUSSIAN ACADEMIC TEXTS

Ivanov V. V. (nomemm@gmail.com)
Innopolis University, Innopolis, Russia

Solnyshkina M. I. (mesoln@yandex.ru),
Solovyev V. D. (maki.solovyev@mail.ru)
Kazan Federal University, Kazan, Russia

This paper addresses the problem of readability assessment for Russian texts and investigates the impact of 24 lexical, syntactic and frequency features. The research was conducted on Russian Readability Corpus containing two sub-corpora, two sets of 5–11 grade level textbooks on Social studies for native speakers of Russian. The sub-corpora were collected for research purposes, annotated and marked as BOG and NIK. The application of the Ridge regression has demonstrated the connection between readability and average sentence length, average number of coordinating chains, average number of sub-trees, frequency and lexical features. The results of the study have the potential to be applied in a wide variety of areas including primarily education, as well as webpage design, document management.

Key words: readability assessment, Russian Readability Corpus, average sentence length, average number of coordinating chains

ЭФФЕКТИВНОСТЬ ПРИЗНАКОВ ДЛЯ АНАЛИЗА СЛОЖНОСТИ АКАДЕМИЧЕСКИХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Иванов В. В. (nomemm@gmail.com)
Университет Иннополис, Иннополис, Россия

Солнышкина М. И. (mesoln@yandex.ru),
Соловьев В. Д. (maki.solovyev@mail.ru)
Казанский федеральный университет, Казань, Россия

В статье рассматривается проблема оценки удобочитаемости для российских текстов и исследуется влияние 24 лексических, синтаксических и частотных признаков. Исследование проводилось на российском корпусе (школьных) учебных текстов, содержащем два набора учебников уровня 5–11 класса по обществознанию для носителей русского языка. Две части корпуса, составляют учебники, написанные двумя разными авторами, были аннотированы и обозначены как BOG и NIK. Применение метода Ридж-регрессии продемонстрировало связь удобочитаемости со средней длиной предложения, средним числом координационных цепочек слов, средним количеством поддеревьев, частотой и лексикой. Результаты исследования могут быть применены в самых разных областях, включая прежде всего образование, а также дизайн веб-страниц, управление документами.

Ключевые слова: оценки удобочитаемости, Russian Readability Corpus, средняя длина предложения

1. Introduction

In the Russian Federation today, educators, parents and administrators are buzzing about Unified National Examinations, which are expected to mark a big shift to better practices of assessment. The latter is impossible if educators are not provided with a wide range of leveled reading materials to tailor all categories of students' learning programs. To achieve desired learning outcomes students and educators need available databases of leveled reading materials and textbooks to match various 'reader—text' profiles'. As for the textbook writers, they are expected to create books tailoring a wide range of abilities and goals but providing a minimal core syllabus for all categories of students (https://russian.rt.com/russia/article/434027-ministr-obrazovaniya-vasileva-intervyu). Special attention should also be paid to profiles of children with specific reading comprehension difficulties (https://alldef.ru/ru/articles/almanah-13/edinaja-koncepcija-specialnogo-federalnogo-gosudarstvennogo).

Unfortunately, the existing textbooks which play the central role in teaching are traditionally criticized for being "nothing but collections of facts" (http://www.nlobooks.ru/node/2808) and for "complicated language" (https://www.znak.com/2014-04-08/pochemu_odin_iz_samyh_populyarnyh_uchebnikov_po_matematike_ne_proshel_gosudarstvennuyu_ekspertizu).

Realizing vital importance of reading for national progress, in 2003 Russia launched a sustainable "The National Program of Support and Development of Reading" which announces that "Modern Russia has approached a critical threshold in its neglect of reading on the national scale and at the moment we witness the beginning of the process of irreversible destruction of the nucleus of national culture" (http://www.library.ru/1/act/doc.php?o_sec=130&o_doc=1122). The program calls for evoking interest of younger generation in reading and turning Russians into "active readers". The Program also specifies the significance of "improving the quality and variety of readable literature in all areas of knowledge" and "establishing a system of selecting books for different categories of readers" (http://www.library.ru/1/act/doc.php?o_sec=130&o_doc=1122).

The research held in 2016–2017 showed that reading comprehension skills of Russian primary schoolchildren aged 9–10 top the list of international ranking, however, by the age of 15 Russian secondary schoolchildren gradually move to the middle of the ranking (http://docs.cntd.ru/document/436739637). All the above makes the problem of finding reading material of the right difficulty and assessing educational text readability relevant and even critical in realizing national goals.

As a part of a bigger research aimed at computing a readability formula for Russian texts, in this paper we address the following research question: what features in a linear regression model are informative for estimating readability of Russian academic texts.

2. Related work

Though studies on assessment of texts readability and readability formulas have a history of over a century [Chall, 1958], the Russian history of estimation of text readability is much shorter. Readability as a quantitative concept and a function of text variables was addressed for the first time as late as in the 1970s and 1980s (Lerner, 1974, Ushakov, 1980, Tomina, 1985, Tsetlin, 1980, Mackovskij 1976). By now Russian text analysts have five readability formulas at their disposal:

- Flesch Reading Ease Readability Formula
  $206.835 - (1.3 \times ASL) - (60.1 \times ASW)$

- Mikk [1970]: $0.01 \times x1 + 0.27 \times x2 + 0.54 \times x3$
- Mackovskij [1976]: $0.62 \times ASL + 0.123 \times X4 + 0.051$
- Tuldava (1975): $i \times lg(j)$,
- Oborneva (2006): $206.836 - (1.52 \times ASL) - (65.14 \times ASW)$

where:

- ASL, j = Average Sentence Length, the number of words divided by the number of sentences)
- S = the average number of sentences per 100 words.
- ASW, i = Average number of syllables per word, the number of syllables divided by the number of words),
- x1 = the length of sentences in the number of printed characters,
- x2 = the percentage of different unfamiliar words,
- x3 = the abstractness of the repetitive notions expressed by nouns,
- X4 = the percentage of more than 3-syllable words.

Though the threshold between short and long sentences or at least the number beyond which readability declines for all readers have never been adequately defined the average sentence and word length have always been viewed as good indicators of readability in the majority of readability formulas for Russian texts (see above).

Current studies on texts readability prove strong relationship between word frequency and text readability and provide concrete options for more effectively making use of lexical frequency information in practice [Chen, X. B. & Meurers, D., 2016]. The results of Russian researchers' studies also show that text readability estimation should take into account the distribution of a range of lexical features in a text [see Mikk, 1970, Sharoff, 2008].

Extensive studies were also conducted on the impact of syntax on readability of Russian texts. As features influencing text readability, include the following: the number of participles, adverbial participles, the number of participial constructions, the number clauses in a complex and compound sentences. The researchers specifically emphasize the importance of different connectives such as conjunctions in compound sentences [Krioni, N. K., Nikin, A. D. & Filippova, A. V., 2008]. Far from being solved, the problem of readability correlation with text syntactic features still remains a challenging and highly relevant research area.

3. Feature analysis and model selection for text complexity prediction

In the model selection our main aim is to define an appropriate subset of features for a linear regression model. As described above such models are well-known and are based on two or three parameters (such as average sentence length or average syllables per word). To the best of our knowledge there was no investigation of a wider set of features for prediction of text complexity in Russian. Given the small number of texts in the corpus, we are focus on those linear models, that will not overfit.

3.1. Description of features

In this paper we explored an extended feature set for text complexity modeling. The first part of the feature set contains features based on length and frequency. This part includes 'average words per sentence', 'average syllables per word' and 'frequency of content words' (FREQ). The FREQ feature is calculated using the Russian

frequency dictionary. We count frequencies for each word in an input text.The second part of feature set includes features calculated from part-of-speech tags. In fact, these features represent number nouns, adjectives, verbs, pronouns and negations occur in a text. The POS-tags were derived using TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/). The third part includes syntactic features derived with ETAP-3 system.

PART1: Features based on length and frequency:

- FREQ is a cumulative frequency of content words,
- ASL is an average number of words per sentence,
- ASW is an average number of syllables per word.

PART2: Features based on POS tags:

- NOUNS is a number of nouns per sentence,
- VERBS is a number of verbs per sentence,
- ADJ is a number of adjectives per sentence,
- PRONOUNS is a number of pronouns per sentence,
- PERONAL PRONOUNS is a number of personal pronouns per sentence,
- NEG is a number of negations per sentence.

PART3: Features based of syntactic dependencies:

- AVERAGE_PATH is the quotient of the number of nodes and the number of leaves in a sentence.
- AVERAGE_SOCHIN_LENGTH is the average length of coordinating constructions
- DEEPRICH_RATE is the average number of verbal participles.
- DEEPRICH_V is the average span of a verbal adverb phrase.
- LEAVES_NUMBER is the average number of 'leaves' (terminal nodes, i.e., words that are not anyone's "hosts") in a sentence.
- LONGEST_PATH is the average length of the longest branch.
- NOUNS_DEP is the average number of modifiers in a nominal group; coordinating and explanatory links are ignored.
- PODCHIN_NUMBER is the ratio of sentences in which there is at least one subordinate conjunctions or relational links.
- PODCHIN_RATE is the average number of subordinate links.
- PRICH_RATE is the average number of participial construction; participial constructions are defined as a participle that has at least one dependent.
- PRICH_V is the average span of a participial construction is the quotient of the number of nodes that depend on the participle.
- SENTSOCH_NUMBER is the average number of compound sentences.
- SOCHIN_NUMBER is defined as the average number of coordinating chains.
- PATH_NUMBER is defined as the average number of sub-trees (in a sentence).
- VERBS_DEP is defined as the average number of finite dependent verbs and is calculated as the sum of nodes directly dependent on the finite verb divided by the number of finite verbs; coordinating and explanatory links were ignored.

3.2. Description of the corpus

The first major statistical issue in building a corpus of texts, as Biber (1990) puts it, "concerns the sampling of texts: how linguistic features are distributed across texts and across registers, and how many texts must be collected for the total corpus and for each register to represent those distributions?" Having compared the internal variations of the two texts in the corpus, Biber (1990) concludes that text samples of 1000 words are representative for the text categories under study. He also proved that the 20–80 samples of texts are enough for correlation-based analysis [Biber 1990].

Two collections of texts were assembled for the research. The first collection of 7 texts derived as a result of OCR and postprocessing of textbooks on Social Studies by L. N. Bogolubov. We mark this collection as "BOG". Textbooks cover range of 6–11 Grade Levels. The second collection of 7 texts from textbooks on Social Studies by A. F. Nikitin marked "NIK" aimed at 5–11 Grade Levels. Further we refer to the two collections collectively as a Russian Readability Corpus (RRC). Both sets of textbooks are from the "Federal List of Textbooks Recommended by the Ministry of Education and Science of the Russian Federation to Use in Secondary and High Schools". To ensure reproducibility of results, we uploaded the corpus on a website (http://kpfu.ru/slozhnost-tekstov-304364.html). Note, however, that the published texts contain shuffled order of sentences. This shuffling, indeed, does not affect the values of features, because they do not depend on sentence order. **Table 1** provide numerical description of the RRC.

Table 1. Numerical Data on RRC

Grade level	Tokens		Sentences		ASL		ASW	
	BOG	NIK	BOG	NIK	BOG	NIK	BOG	NIK
5-th	—	17,221	—	1,499	—	11.49	—	2.35
6-th	16,467	16,475	1,273	1,197	12.94	13.76	2.56	2.71
7-th	23,069	22,924	1,671	1,675	13.81	13.69	2.84	2.70
8-th	49,796	40,053	3,181	2,889	15.65	13.86	2.96	2.88
9-th	42,305	43,404	2,584	2,792	16.37	15.55	3.04	3.00
10-th	75,182	39,183	4,468	2,468	16.83	15.88	3.07	3.12
10-th*	98,034	—	5,798	—	16.91	—	3.05	—
11-th	—	38,869	—	2,270	—	17.12	—	3.11
11-th*	100,800	—	6,004	—	16.79	—	3.19	—

Comment. Star sign (*) denotes advanced versions of books for the corresponding grade; sign '—' denotes absence of a textbook for the corresponding grade.

3.3. Analysis of features

3.3.1. Correlation between features

We provide the results of correlation analysis in the **Table 2**. In general, some syntactic features are similar to others and correlate with the target variable (readability,

measured as a grade level). However, it is evident that all the syntactic features have lower correlation coefficient with the target feature ('Grade Level'), than the two 'classical' lexical features (ASL and ASW) do.

Table 2. Correlation between features and target feature, grade level

	Feature name	Correlation coefficient		Feature name	Correlation coefficient
1	ASL	0.94	13	NOUNS	0.82
2	ASW	0.94	14	VERBS	0.74
3	SOCHIN_NUMBER	0.93	15	NEGATIONS	0.70
4	PRICH_RATE	0.91	16	PRONOUNS	0.70
5	NOUNS_DEP	0.88	17	PODCHIN_RATE	0.64
6	AVERAGE_ SOCHIN_LENGTH	0.87	18	PODCHIN_NUMBER	0.62
			19	DEEPRICH_V	0.52
7	PATH_NUMBER	0.87	20	PERS_PRONOUNS	0.47
8	LONGEST_PATH	0.84	21	DEEPRICH_RATE	0.44
9	FREQ	0.84	22	VERBS_DEP	0.43
10	LEAVES_NUMBER	0.84	23	PRICH_V	0.33
11	AVERAGE_PATH	0.84	24	SENTSOCH_ NUMBER	0.03
12	ADJ	0.82			

3.3.2. Significance of features

We tested significance of a linear regression model in the following setting. We applied the F-test for linear regression to evaluate whether any of the independent variables in a multiple linear regression model are significant. The results of the F-test are presented in the table below. P-values are denoted with '**' and '*' signs.

Table 3. Results of F-test for significance of attributes of a linear regression model (** corresponds to p-values < 0.01; * corresponds to p-values < 0.05)

	Feature name	F-score		Feature name	F-score
1	ASL	95.58**	13	VERBS	24.49**
2	ASW	91.93**	14	NOUNS	19.17**
3	SOCHIN_NUMBER	71.23**	15	NEGATIONS	14.11**
4	PRICH_RATE	56.20**	16	PERS_PRONOUNS	11.00**
5	NOUNS_DEP	42.17**	17	PODCHIN_RATE	8.35*
6	AVERAGE_ SOCHIN_LENGTH	38.91**	18	PODCHIN_NUMBER	7.41*
			19	DEEPRICH_V	4.49
7	PATH_NUMBER	35.69**	20	DEEPRICH_RATE	2.86
8	LONGEST_PATH	29.45**	21	VERBS_DEP	2.76
9	FREQ	29.32**	22	PRICH_V	1.42
10	LEAVES_NUMBER	29.01**	23	PRONOUNS	0.22
11	AVERAGE_PATH	28.60**	24	SENTSOCH_ NUMBER	0.01
12	ADJ	25.33**			

It was expected that, that the most significant attributes include well-known features on length of sentences and words (ASL, ASW), syntactic features (such as SOCHIN_NUMBER, PRICH_RATE, etc.) and lexical attributes (ADJ, VERBS, etc.).On the other hand, insignificant features include SENTSOCH_NUMBER, PRICH_V, etc. which corresponds to correlation analysis. We use results of this evaluation for filtering insignificant features. Therefore, based on p-value (<0.01), for further analysis we keep only first 16 features from the **Table 3**. It is clear that 16 features are too many to build a robust linear regression given the number of texts in our corpus.

In the next step we make use a technique for feature selection: Ridge regression [Wessel N. van Wieringen, 2018] to find a subset of most relevant features for a prediction model. An alternative is just a brute-force search for the best subset of features. A drawback of thebrute-force approach is clear: given the number of texts in the corpus a model with many featurescan easily overfit the data even if we split the dataset into a train and test sets.

3.3.3. Feature selection with Ridge regression

Ridge regression is an approach that represents regularization technique with constrain (L2-norm) on the feature weights in a linear model. The approach can be used to rank features with respect to their magnitude (their influence on the target variable). We use the ranked list of features to select reasonable subset of features for linear regression model of text complexity.

Table 4. Ridge regression results in feature selection

	Feature	Absolute value of Coefficient in Ridge Regression		Feature	Absolute value of Coefficient in Ridge Regression
1	**ASL**	0.506	9	NOUNS_DEP	0.071
2	ASW	0.125	10	FREQ	0.034
3	SOCHIN_NUMBER	0.119	11	NEGATIONS	0.010
4	PRICH_RATE	0.106	12	AVERAGE_PATH	0.007
5	LONGEST_PATH	0.089	13	PERS_PRONOUNS	0.003
6	PATH_NUMBER	0.079	14	VERBS	0.001
7	LEAVES_NUMBER	0.075	15	ADJ	0.001
8	AVERAGE_ SOCHIN_LEN	0.071	16	NOUNS	0.000

4. Discussion of results and conclusion

With the view of increasing amount of available academic texts, broadening varieties of alternative training options and personalized training, the problem of selecting appropriate teaching materials is becoming urgent. Textbooks of almost the same content may differ in the degree of complexity (readability) of presentation. To the best of our knowledge, there have been no extensive multi-feature studies

of readability of Russian texts. The authors of the paper offer an innovative 24-feature analysis of Russian texts readability embracing "classical" frequency features, part-of-speech, and syntactic features. For our research we create dataset which are uploaded on KFU website and are available for potential verification and validation of the research outcomes.

The results derived in this paper support the following points. First, average sentence length is the most important feature for text complexity prediction. Second, there are several highly important syntactic features such as the average number of coordinating chains, average number of sub-trees, as well as frequency and lexical features that can improve prediction. Third, surprisingly, average syllables per word may not be a very important feature (in presence of other features), even though it correlates with target variable.

The results obtained in this article are far from being final, since they are received on a relatively small corpus of homogeneous texts. Readability of different types of texts is to be estimated with different formulas. Rather, this article offers a methodology for this type of research. We intend to further apply the proposed approach to texts of other subject areas and genres. It is also proposed to further expand the set of text features to be studied, including semantic and discursive features. The research available suggest that lexical features of reading texts such as word frequency, word identification ability, mean noun frequency level as well as lexical diversity and type-token ratio (TTR) are factors that influence reading comprehension, it is this fact that makes them reliable metrics in assessing text complexity [see Solovyev 2018]. Based on the hypothesis that the average word frequency across the textbooks is to have consistent progression, we plan to conduct a cross-sectional (grade) study of textbooks for different age groups with regard to lexical density, TTR and lexical diversity.

Though selecting appropriate reading text for students of different grades is of crucial importance, we do not narrow our studies to educational needs only. The research shines a light on issues worthy of discussion with regard to texts used in mass media, healthcare, document management, etc. Contributing this article we hope to attract attention of scholars working in related areas so that we could combine our efforts and change the opportunities for thousands of struggling readers. National discussions are needed to ensure that writers (textbook authors, speech and news writers, journalists, etc.) can make informed decisions about the difficulty level of the texts they generate.

Acknowledgements

This research was financially supported by the Russian Science Foundation, grant № 18-18-00436, the Russian Government Program of Competitive Growth of Kazan Federal University, and the subsidy for the state assignment in the sphere of scientific activity, grant agreement № 34.5517.2017/6.7. The Russian Academic Corpus (section 3.2 in the paper) was created without supporting by the Russian Science Foundation. We would like to convey our sincere gratitude to Dr. Lana Timoshenko and Dr. Ivan Rygaev for their valuable assistance with syntactic annotation of the corpus.

References

1. *Biber, D.* (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. Literary and Linguistic Computing, Literary and Linguistic Computing, 5: 257–69.
2. *Chall, J. S.* (1958). Readability: An appraisal of research and application. Bureau of Educational Research Monographs, No. 34. Columbus, Ohio State Univ. Press.
3. *Chen, X. B., Meurers, D.* (2016). Characterizing Text Difficulty with Word Frequencies. In Proceedings of The 11th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 84–94). San Diego, CA. Association for Computational Linguistics.
4. *Choldin, M. T.* (1979). Rubakin, Nikolai Aleksandrovic. In A. Kent, H. Lancour, J. E. Daily (Ed.) Encyclopedia of library and information science. (pp. 178–179). Basel: CRC Press.
5. *Krioni N. K., Nikin A. D., Filippova A. V.* (2008). Avtomatizirovannaya sistema analiza parametrov slozhnosti uchebnogo teksta In Tekhnologiya i organizatsiya obucheniya : nauch. izdanie. — Ufa: UGATU, P. 155–161.
6. *Lerner, I. Ya.* (1974) Kriterii slozhnosti nekotorykh elementov uchebnika: Problemyshkol'nogouchebnika [The criteria for the complexity of some elements of the textbook: Problems of a school textbook]. Is. 1. Moscow: Prosveshchenie.
7. *Mackovskij, M. S.* (1976). Problemy chitabel'nosti pechatnogo materiala. Smyslovoe vospriyatie rechevogo soobshcheniya v usloviyah massovoj kommunikacii [Problems of readability of printed material. Semantic perception of speech messages in conditions of mass communication]. Moskow: Nauka.
8. *Mikk, Y. A.* (1970) O factorakh ponyatnosti uchebnogo teksta [On factors of comprehensibility of educational texts]. Diss. ... cand. ped. sciences. Tartu.
9. *Mutt, O.* (1984). Aktual'nye voprosy otbora uchebnogo materiala dlya vuzovskogo kursa inostrannogo yazyka [Actual questions of selection of the educational material for the university course of a foreign language]. Tartu: The Tartu State University
10. *Oborneva, I. V.* (2006) Avtomatizirovannaya otsenka slozhnostiuchebnykh tekstov na osnove statisticheskikh parametrov [Automated estimation of complexity of educational texts on the basis of statistical parameters]. Pedagogy Cand. Diss. Moscow.
11. *Sharoff, S., Kurella, S. and Hartley, A.* (2008). Seeking needles in the web's haystack: Finding texts suitable for language learners. In Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8).
12. *Solovyev, V., Ivanov, V., Solnyshkina M.* (2018) Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics. Journal of Intelligent & Fuzzy Systems, (Preprint):1–10, 2018.
13. *Tomina, Yu. A.* (1985) Ob"ektivnaya otsenka yazykovoy trudnosti tekstov (opisanie, povestvovanie, rassuzhdenie, dokazatel'stvo) [An objective assessment of language difficulties of texts (description, narration, reasoning, proof)]. Abstract of Pedagogy Cand. Diss. Moscow.
14. *Tsetlin, B. C.* (1980) Didakticheskie trebovaniya k kriteriyam slozhnosti uchebnogo materiala [Didactic requirements to the complexity criteria of educational material]. Novye issledovaniya v pedagogicheskikh naukakh. 1 (35). pp. 30–33.

15. *Tuldava, Yu. A.* (1975) Ob izmerenii trudnosti tekstov [On measuring the complexity of the text]. Uchenye zapiski Tartuskogo universiteta. Trudy po metodike prepodavaniya inostrannykh yazykov. 345. pp. 102–120.
16. *Ushakov, K. M.* (1980) O kriteriyakh slozhnosti uchebnogo materiala shkol'nykh predmetov [On the criteria of complexity of teaching material of school subjects]. Novye issledovaniya v pedagogicheskikh naukakh. 2 (36). pp. 33–35.
17. *Wessel N. van Wieringen* (2018) Lecture notes on ridge regression. arXiv:1509. 09169v2 [stat.ME] 6 Jan 2018.