Complexity of Russian Academic Texts as the Function of Syntactic Parameters

Valery Solovyev[1], Marina Solnyshkina[1], Vladimir Ivanov[2], and Svetlana Timoshenko[3]

[1] Kazan Federal University, Kazan, Russia
[2] Innopolis University, Innopolis, Russia
[3] Institute for Information Transmission Problems, Moscow, Russia

Abstract. Providing students with academic materials of steadily increasing complexity is equally critical for textbook publishers, teachers, and test developers and automation of text complexity assessment is relevant in a number of areas. The research suggests that the existing automated analyzers of Russian texts have two main limitations: (1) a narrow range of variables which include nothing but word length, sentence length and word frequency; (2) the constants of Russian readability formulas once calculated for fiction texts are being applied to texts of different types and genre. The authors propose an innovative approach addressing limitations of Russian text complexity tools. The authors original algorithm of designing a predictive model of text complexity is based on a training text corpus and a set of syntactic features. The syntactic features calculated by means of ETAP, a system is based on a detailed description of Russian grammar. A linear model for text complexity assessment was evaluated on a corpus of Russian academic texts. We show that syntactic features improve the quality of the model.

1 Introduction

Effective reading comprehension implies that reading materials correspond readers cognitive and language abilities. The idea behind the existing practice in education is to ensure that students are exposed to the age-appropriate materials which are neither too complicated nor too simple for a reader. The "age-appropriateness" has been traditionally measured by the Grade level which is viewed as "what all students need to know and be able to do at each grade level" to progress through their education[4].

Grade level descriptors identify the specific content (knowledge, skills, abilities) and the language of a particular course which students at a particular education stage (or grade) are exposed to[5].

[4] www.k12.wa.us/CurriculumInstruct/learningstandards.aspx

[5] https://www.gov.uk/government/publications/grade-descriptors-for-gcses-graded-9-to-1/grade-descriptors-for-gcses-graded-9-to-1-english-language;

There are also a number of English text complexity analyzers available online for any educator selecting a text for students[6]. The existing automatic analyzers use hundreds of parameters ranging from quantitative, i.e. word length and sentence length only, to qualitative (levels of meaning or purpose; structure; language conventionality and clarity; and knowledge demands) to match a particular reader and a text[7].

T.E.R.A., for instance, is an engine developed in 2012 by SoLET Lab which analyzes five textual components, such as narrativity, syntactic simplicity, word concreteness, referential cohesion and deep cohesion[8]. T.E.R.A. also estimates the grade level of the text using the Flesch-Kincaid Grade Level readability formula [3]. Knowing the grade level of texts in a corpus, educators select the most suitable text for the target audience.

The situation in Russia is different. Significant gaps have been reported between the complexity levels of texts that students are asked to read in high school as well as at tertiary levels and students abilities: the books are either too simple or too complicated for students[9].

Researchers also have evidence of students lack of wish to read[10] who in many cases are caused by the inappropriate selection of a book by an educator. Eliminating the gap between critically important texts and students abilities scholars have been developing tools to profile texts that students would be able to and want to read[11]. Unfortunately, Russian text complexity analyzers so far apply no other variables but quantitative, i.e. word length and sentence length [4]. In the paper we aim at the following research question: Which syntactic text features better correlate with complexity of Russian academic texts?

https://www.discoveractaspire.org/wp-content/uploads/2016/06/grade-5-pld-reading.pdf;
http://oh.portal.airast.org/core/fileparse.php/3094/urlt/OCBA_G4_SocialStudies_PLDs.pdf;
http://www.corestandards.org/assets/Appendix_A.pdf

[6] http://www.readabilityformulas.com/free-readability-formula-tests.php,
https://www.wikihow.com/Determine-the-Reading-Level-of-a-Book

[7] https://readable.io/text/,
http://casemed.case.edu/cpcpold/students/module4/Word_Readability.pdf

[8] http://129.219.222.66/Publish/tera.html

[9] https://elibrary.ru/item.asp?id=20134801,
http://www.miep.edu.ru/uploaded/zvezdova_oreshkin.pdf,
http://www.mdk-arbat.ru/bookcard?book_id=934790

[10] https://www.science-education.ru/ru/article/view?id=22229,
https://cyberleninka.ru/article/v/chtenie-kak-sotsialnaya-problema,
https://trv-science.ru/2015/11/17/pochemu-ne-chitayut-shkolniki/

[11] http://readability.io/,
https://netpeak.net/ru/blog/readability/

2 Background

Statistical linguistics is based on the assumption that a limited number of quantitative characteristics and functional relationships between them, obtained for a limited set of texts, characterize the language as a whole and its functional styles (mass media, academic, research etc.). The accumulated data are used not only to decipher historical letters and reveal peculiarities of individual styles but also to describe types of texts and assess their complexity. The latter is very much demanded in education for text experts, teaching materials developers and test designers. Though the terms ‘text complexity’, ‘text difficulty’ and ‘text readability’ are still sometimes used interchangeably, the notions of the concepts began being separated in Russian academic literature as early as in 1970-s. I.Lerner defined complexity “as a category that characterizes the range of activities necessary to solve a cognitive task, regardless of who performs this activity”. While text difficulty is viewed by the researcher as “a category characterizing a persons readiness to overcome obstacles while comprehending a reading text”.

Accepting I.Lerners point of view V. Tcetlin specified that “complexity of any educational material is its objective characteristics whereas difficulty is a subjective factor of students’ preparedness to overcome complexity”. The general approach to “text readability” once proposed by M. Vogel and K. Washburn was taken as a basis in all subsequent works. It is based on the objective characteristics of the text highly correlating with the quantitative results of a test, and implies designing a regression equation between the success of a text comprehension, on the one hand, and text parameters – on the other [5]. There are also a number of graphic parameters of a text influencing its comprehension such as fonts, indentations, spaces, colors etc., which are beyond the authors interest in the article. The first attempts to assess Russian texts complexity were made in late 1970-s, about 50 years later than the corresponding English studies [2].

In 1985 Yu. Tomina suggested that lexical indicators of the linguistic difficulty of texts are the number of unfamiliar words and abstract words, while syntactic indicators are the number of participial constructions, the number of similar parts of a sentence, prepositional-nominal groups[12]. The formulas proposed to predict Russian texts readability were based on a number of objective variables borrowed from the similar English texts complexity formulas, i.e. word length and sentence length. Assessments of texts readability were initially carried out by hand, and then in 2010s – by means of computer programs. They were all based on I. Oborneva’s readability formula derived in 2006 [4]:

$$FRE = 206.836 - (1.52 \cdot ASL) - (65,14 \cdot ASW)$$

.

Here, ASL is the average sentence length, i.e., the number of words divided by the number of sentences; ASW is the average number of syllables per word, i.e., the number of syllables divided by the number of words in a text. The constants

[12] http://www.dissercat.com/content/obektivnaya-otsenka-yazykovoi-trudnosti-tekstov-opisanie-povestvovanie-rassuzhdenie-dokazate#ixzz53nzDNXFp

were calculated based on the similar English formulas as well as the comparison of 100 parallel English/Russian literary texts and words in two academic dictionaries Slovar russkogo yazyka pod redaktsyey Ozhegova with 39174 words and Muller English-Russian Dictionary with 41977 words.

3 Datasets

Two collections of texts were assembled for the research. The first collection of 7 texts from textbooks on Social Studies by L. N. Bogolubov marked "BOG" was selected to teach the predictive model and define independent variables of the text variation in the range of 5 – 11 Grade Levels. The second collection of 7 texts from textbooks on Social Studies by A.F. Nikitin marked "NIK" also aimed for 5 – 11 Grade Levels. Further we refer to the two collection 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"[13].

To ensure reproducibility of results, we uploaded the corpus on a website thus providing its availability online[14]. Note, however, that the published texts contain shuffled order of sentences. The sizes of BOG and NIK collections of texts are presented in Table 1.

Table 1. Properties of the preprocessed corpus.

	Tokens		Sentences		Words per sentence		Syllables per word	
Grade	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	–

4 Methods

4.1 Corpus preprocessing

For the sake of convenience, we have preprocessed all texts from the corpus in the same way. Common preprocessing included tokenization, splitting text into sentences and part-of-speech tagging (using the TreeTagger for Russian[15]).

[13] http://www.fpu.edu.ru/fpu/

[14] this link removed by authors due to the blind review process

[15] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

During the preprocessing step we excluded all extremely long sentences (longer than 120 words) as well as too short sentences (shorter than 5 words) which we consider outliers. Clearly, such sentences can be not outliers at all in another domain, but in case of school textbooks on Social Studies sentences shorter than 5 words are outliers.

Extremely short sentences mostly appear as names of chapters and sections of the books or as a result of incorrect sentence splitting. We omit those sentences, because the average sentence length is a very important feature in text complexity assessment and hence should not be biased due to splitting errors. At the same time sentences with five to seven words in Russian can still be viewed as short sentences.

4.2 Lexical level Features

We have explored an extended feature set for text complexity modeling:

- frequency of content words (FREQ),
- average words per sentence (ASL),
- average syllables per word (ASW), and
- features based on POS-tags:
  - number of nouns per sentence (NOUNS),
  - number of verbs per sentence (VERBS),
  - number of adjectives per sentence (ADJ),
  - number of pronouns per sentence (PRONOUNS),
  - number of personal pronouns per sentence (PERS. PRONOUNS),
  - number of negations per sentence (NEG),
  - number of connectives per sentence (CONN).

We have tested the features for their importance in linear regression model. The two features have shown better performance than others: FREQ and ADJ. To assess the quality of proposed models the mean squared error (MSE) and the coefficient of determination R^2 were used. For the results of fitting the parameter values in the corpus, see Table 2.

Table 2. Results of fitting 3-parameter linear model on the RRC dataset.

Features	Formula for grade level	R^2	MSE
ASL, ASW, FREQ	1.92 + 0.5 x ASL + 2.12 x ASW -2.36 x FREQ	0.94	0.22
ASL, ASW, ADJ	1.59 + 0.23 x ASL + -1.48 x ASW 3.97 x ADJ	0.95	0.19

4.3 Syntactic level features

The modern tools for processing of Russian texts are able to extract several syntactic features of texts thus permitting to include a number of syntactic structures as features in readability formulas. To this end we use a syntactic analyzer "ETAP-3" that employs a very detailed description of Russian grammar.

The parser of the multipurpose linguistic processor ETAP-3 is a program that performs parsing. In linguistic terms, parsing results in a dependency tree structure, the nodes of which are word tokens of the input sentence, and 'the edges' are the established syntactic dependencies. Thus, every tree node corresponds to a word token in the sentence processed, whilst the directed arcs are labeled with names of syntactic relations [1]. All the syntactic dependencies have directions, therefore all the dependencies have the original node, its host, and the final one - the dependent node. Each word token is represented in the form of the initial form of a word and a set of its morphological characteristics [1]. All the texts in the collection were processed with the syntactic parser: each sentence was converted into a dependency tree structure. Then the following which the following 14 numeric features were extracted:

- An **'average_path'** is the quotient of the number of nodes and the number of leaves in a sentence
- An **'average_sochin_length** as the average length of coordinating constructions is the number of nodes in branches starting with coordinating constructions divided by the number of such branches; all types of nodes are processed including conjunctions and modifiers
- The **'deeprich_rate** as the average number of verbal participles (verbal adverb phrases) is the quotient of the number of verbal participles and the number of sentences. The verbal participles are defined as a verbal adverb with at least one dependent modifier.
- The **'deeprich_v** as the average span of a verbal adverb phrase is the number of verbal adverb dependent nodes (in all "branches") divided by the number of verbal adverb phrases.
- The **leaves_number** or the average number of "leaves" (terminal nodes, i.e., words that are not anyone's "hosts") in a sentence. Calculation formula: the number of all "leaves" in the text is divided by the number of sentences.
- The **longest_path** as the average average length of the longest branch is the sum of the lengths of the longest branches of the sentence divided by the number of sentences.
- The **nouns_dep** as the average number of modifiers in a nominal group, i.e. the sum of the all the nodes that depend on the nouns divided by the number of nouns; coordinating and explanatory links were ignored.
- The **podchin_number**, i.e. the ratio of sentences in which there is at least one syndetic with subordinate conjunctions or relational links calculated as the number of sentences with at least one link of the type divided by the number of sentences.
- The **podchin_rate**, i.e. the average number of subordinate links calculated as the number of syndetic with subordinate conjunctions and relational links divided by the number of sentences
- The **prich_rate** as the average number of participial construction is calculated as the number of participial constructions divided by the number of sentences; participial constructions are defined as a participle that has at least one dependent.

- The **prich_v** as the average span of a participial construction is the quotient of the number of nodes that depend on the participle (in all "branches") and the number of participial constructions.
- The **sentsoch_number** as the average number of compound sentences is the quotient of the number of coordinating constructions and the number of sentences.
- The **sochin_number** is defined as the average number of coordinating chains and calculated by dividing the average number of coordinating chains in the sentence by the number of sentences; a chain is a sequence of nodes connected by the "coordinating" links, thus conjunctions "break" the chain.
- The **path_number** is defined as the average number of sub-trees (in a sentence), calculated with an external algorithm.
- The **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.

The list of features extracted by the ETAP syntax parser was preprocessed to group similar features. We provide the results of correlation analysis in the Table 3. In general, some syntactic feature are similar to others and correlate with the target variable (readability, measured as a class number). 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 3. Correlation between features and the 'Grade Level'.

Feature name	Correlation coefficient with the 'Grade Level'
ASL	0.94
ASW	0.94
sochin_number (SN)	0.93
prich_rate (PR)	0.91
nouns_dep (ND)	0.88
average_sochin_length (SL)	0.87
path_number (PN)	0.87
longest_path (LP)	0.84
leaves_number (LN)	0.84
average_path (AP)	0.84
podchin_rate	0.64
podchin_number	0.62
deeprich_v	0.52
deeprich_rate	0.44
verbs_dep	0.43
prich_v	0.33
sentsoch_number	0.03

Nevertheless, syntactic features have high correlation with the target variable. This information could be useful for readability and text complexity prediction in Russian. We evaluate syntactic features in the next subsection with respect to the capability to serve as predictors in a linear regression model. Further we consider features with correlation coefficient above a certain threshold (above the line as depicted in Table 3).

4.4 Evaluation of syntactic features

For evaluation of the syntactic features we carried out the two experiments. In the first experiment, we clustered the syntactic features with respect to their similarity to each other. By similarity we treat the correlation of the features' values derived in RRC. The derived groups of features are the following:

- Group 1, (G-1): Features related to the structure of the syntax tree
  - leaves_number (LN)
  - average_path (AP)
  - longest_path (LP)
  - path_number (PN)
- Group 2, (G-2): Features related to noun and participial linear sequences
  - prich_rate (PR)
  - nouns_dep (ND)
- Group 3, (G-3): Features related to coordinating constructions
  - average_sochin_length (SL)
  - sochin_number (SN)

The selected syntactic features could serve better predictors in a linear regression model due to their high correlation with the target variable. We measured the performance of the resulting model in the following way. A linear regression model was trained on the 'BOG' collection and tested on the 'NIK' collection and vice versa. In both cases we use the MSE as a measure of model's performance 4. First, we evaluate three features with highest correlation: SN, PR and ND. Next, three rows of the Table 4 correspond to the groups of syntactic we found. Finally, we evaluate three features, one coming from a certain group. From each group we pick a feature with highest correlation: PN from G-1, PR from G-2 and SN from G-3. It can be seen form the Table 4 that syntactic features from groups G-2 and G-3 are better than those from G-1.

Table 4. Results of syntactic features evaluation in linear models.

Syntactic Features	MSE on 'BOG'	MSE on 'NIK'
SN (G-3), PR (G-2), ND (G-2)	1.15	2.11
Group 1	7.74	25.42
Group 2	1.34	1.32
Group 3	0.55	2.58
PN (G-1), PR (G-2), SN (G-3)	3.2	5.39

In the second experiment, we compared syntactic features to the lexical level features (ASW, ASL, FREQ and ADJ) with respect to their performance. Finally, we have built linear models using combinations of both syntactic and lexical features (Table 5). Syntactic features without any other features leads to poor results. However, combination of lexical level features (ASL and ASW) with syntactic features improves performance of linear model for readability assessment.

Table 5. Results of linear models performance.

Feature set	MSE on 'BOG'	MSE on 'NIK'
ASL, ASW	0.44	0.76
ASL, ASW, FREQ	1.97	0.39
ASL, ASW, ADJ	**0.26**	1.02
ASL, ASW, ALL_SYNTAX_FEATURES	1.18	1.94
ALL_SYNTAX_FEATURES	2.77	6.97
G-1, G-2, G-3	2.6	2.28
ASL, ASW, G-1, G-2, G-3	2.11	**0.29**

5 Discussion and Future work

The main challenge to cope with is the selection of optimal values for the constants in the formulas. We show that syntactic features could be useful in readability assessment. In future research we plan to apply semantic features, such as features based on syntactic n-grams [7, 6] and other types of information extracted from text [8].

Acknowledgements

This research was supported by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, agreement #34.5517.2017/6.7.

References

1. I. Boguslavsky, L. Iomdin, A. Lazursky, L. Mityushin, V. Sizov, L. Kreydlin, and A. Berdichevsky. Interactive resolution of intrinsic and translational ambiguity in a machine translation system. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 388–399, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
2. W. H. DuBay. The principles of readability. *Online Submission*, 2004.

3. J. Kincaid, R. Fishburne Jr, R. Rogers, and B. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.
4. I. Obobroneva. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov [semiautomatic evaluation of the complexity of academic texts on the base of statistic parameters]. moscow: Rae institute of content and methods of teaching. *M.: RAS Institut soderzhaniya i metodov obucheniya*, 2006.
5. Y. Shpakovskiy et al. *Otsenka trudnosti vospriyatiya i optimizatsiya slozhnosti uchebnogo teksta [Evaluation of the difficulty of perception and optimization of the text complexity].* PhD thesis, 2007.
6. G. Sidorov. Non-linear construction of n-grams in computational linguistics. *México: Sociedad Mexicana de Inteligencia Artificial*, 2013.
7. G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández. Syntactic dependency-based n-grams as classification features. In *Mexican International Conference on Artificial Intelligence*, pages 1–11. Springer, 2012.
8. V. Solovyev and V. Ivanov. Knowledge-driven event extraction in Russian: corpus-based linguistic resources. *Computational intelligence and neuroscience*, 2016:16, 2016.