Characterizing Text Complexity with Core Vocabulary Distributional Patterns: Corpus-based Approach

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

[1] Kazan Federal University, 18, Kremlyovskaya st., Kazan, Russia
maki.solovyev@mail.ru, mesoln@yandex.ru
[2] Innopolis University, 1, Universitetskaya st., Innopolis, Russia
v.ivanov@innopolis.ru

Abstract. In this paper, we report a corpus study aimed at testing the hypothesis that bigram distributional information is related to text complexity. We explored a corpus of Russian textbooks on Social studies for middle and high school to examine how the number of bigrams of the core vocabulary correlates with reading levels of texts within the grade range 5 – 11. The corpus contains 45380 sentences from 14 textbooks, written by two independent groups of authors. Each word in the corpus has a part-of-speech tag derived by TreeTagger. Due to the nature of the domain, we focus our study on a single, but high-frequency pattern: *'chelovek' (a man) + verb*. The findings are particularly relevant for text complexity theory as they are consistent with the previous results of corpus investigations on a correlation of text complexity with a number of text features.

Keywords: a corpus, bigram, text complexity, distributional patterns

1 Introduction

The research presented in this article is a part of Russian Academic Text Complexity (RATC) project, which aims at defining roles of different metrics in academic text complexity analysis and has been carried out at Kazan Federal University for over a year. The ultimate goal of the project is to provide cognitive and linguistic profiles of Russian academic texts for middle and high schools based on the complex linguistic analyses of the latter, describe and determine correlations between text complexity (or grade levels) and academic text features [18,9]. This current study is aimed at exploring two research questions: 1. To what extent the number of collocates (and ngrams) of a particular noun may constitute a valid and reliable index that can be used to objectively discriminate between texts of different grade levels? 2. How does the size of a semantic class of verbs correlate with text complexity across the grade levels 5 – 11?

2 Related works: Lexical Features in Text Complexity Studies

In modern text complexity studies, lexico-semantic features of reading texts are considered valuable and essential metrics in assessing text complexity (see [18]). The research shows that word frequency as a text feature impacts accuracy of perception [8], word identification ability of readers [15] and readers' speed of performance in language tasks [13].

At present, lexical metrics are used in over one hundred readability formulas including those of Spache [19] and Dale [3] (see [10,7]). In 1969, W.B. Elley suggested using the term and feature of "mean noun frequency level" to define readability levels of texts [5]. Another reliable feature discriminating text readability and its grade profile is found to be text lexical diversity (LD) or variation defined as 'the range and variety of vocabulary deployed in a text by either a speaker or a writer' [14]. The type-token ratio (TTR), i.e. the number of word types (or different words) divided by the number of tokens.

Later on, TTR was acknowledged to be sensitive to the text length, and numerous revised indices such as Root TTR and Corrected TTR, which take the logarithm and square root of the text length instead of the direct word count as denominator were suggested and proved to produce better results [22]. Experts in Text complexity also admit that "a phrase (n-gram) gives more information than just a single word" and better represents a text than just a word [2] as it provides better document representation than simple "Bag of Words". Semantic classes of words sharing a number of meaning components are viewed in Natural Language Processing as classes not only useful for predicting certain correlations between syntax and semantics Another text feature, i.e. lexical tightness, which is proved to strongly correlate with grade level in "a collection of expertly rated reading materials" (p.29) is viewed by M.Flor et al. [6] as a metric representing the degree to which a text tends to use words that are highly inter-associated, i.e. semantically connected in the language. All these provide a foundation for the hypothesis that the number of bigrams (and semantic connections) of a high-frequency word in a text has a tendency to grow alongside with text complexity thus representing text complexity. In present study we introduce text corpus of academic texts and use it to investigate the abovementioned hypothesis.

3 Corpus Description

For the purpose of this study, Russian Academic Corpus (RAC) was compiled of two batteries of textbooks for Russian students: edited by Bogolyubov and by Nikitin. In the Russian Federation the course on Social Studies is taught for 7 years: it starts in Grade 5 when children are typically aged 11 and finishes in Grade 11 where the predominant majority of students is 17 years old. The course finishes with a matriculation exam in the 11th Grade when a certain number of students select the subject for a high-stake exam to continue their education at universities.

To ensure reproducibility of results, we uploaded the corpus on a website thus providing its availability online[3]. 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		ASL		ASW	
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	–

3.1 Preprocessing of the corpus

For the convenience, we have preprocessed all texts from the corpus in the same way. Common preprocessing included tokenization and splitting text into sentences. During the preprocessing step we excluded all extremely long sentences (longer than 120 words[4]) 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 for the case of school textbooks on Social Studies sentences shorter than 5 words are outliers. Sentence and word-level properties of the preprocessed dataset are presented in Table 1.

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, because the average sentence length (in our corpus) is higher than ten.

The last two columns in Table 1 present well-known features that have been widely exploited for assessment of readability of English texts. Average sentence length (or average words per sentence, ASL) and average syllables per word[5],

[3] http://kpfu.ru/portal/docs/F1554781210/shuffled.zip

[4] Indeed, very long sentences appear as long citations from official documents which styles are completely differ from school textbooks

[5] Number of syllables in a Russian word can be computed as a number of vowels in the word.

ASW, are the parameters in Flesch and Flesch-Kincaid formulas [18]. Table 1 demonstrates that values of ASL and ASW, as it is generally expected, increase with the grades.

All annotations in the corpus are performed on three levels: text-level, sentence-level and word-level. At the text-level meta-annotations refer to a number of sentences and a set of tokens[6], an author and a grade-level of a given text.

3.2 Selection of pattern for analysis

At the word-level we have part-of-speech tag for each word. POS-tagging has been performed with the use of the TreeTagger for Russian[7]. The tagset is available from the website of the project. As reported in [4] accuracy of 96% on POS tags and of 92% on the whole tagset was achieved by TreeTagger for Russian. Kuzmenko [11] reported POS tagging accuracy of TreeTagger between 88% and 95% depending on dataset.

The contrastive analysis proved high frequency of content words, mostly nouns and adjectives thus confirming two texts characteristics: (1) all the texts studied are qualified as informative; (2) their narrativity level is much lower than that of fiction texts [17]. The research also identified the word 'chelovek'(a man) to be the most frequent noun in the corpus. The word (in different forms) occurs 7447 times which is around 3.1% of all nouns' mentions. In the list of most frequent nouns it is followed by nouns such as 'law' ('pravo'), 'society' ('obschestvo'), 'life' ('zhizn').

4 Analysis of collocations in corpus

4.1 Collocations as a Feature of Text

Corpus studies reveal patterns of word use in natural languages [Hanks 2004]. These patterns can be analyzed and applied in text complexity research as a metric which shows semantic distinctions of words in contexts of different complexity. We computed the Corpus with the aim to retrieve the list of verbs with which the word 'chelovek' (a man) collocates and find out how the semantic range and variety of these verbs extend across the textbooks of grade levels 5 – 11. The semantic range of verbs was analysed based on based on "Russian semantic dictionary" [16] in which the author provides a taxonomy of 35000 verbs and according to their meanings classifies them into 3 broad and numerous fine-grained subclasses. Shvedova's taxonomy as well as all traditional semantic classifications (see [12,21,1]) goes back to the idea of semantic fields of Trier [20] and defines three main classes of verbs:

- functional are the verbs with weakened and/or incomplete content meaning: link-verbs and semi-notional verbs joining subjects and predicatives, verbs

[6] Tokens include words, numbers, punctuation, etc.
[7] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

denoting the Beginning, Middle and End of an Event, modal verbs, verbs denoting connections, relations and naming, deictic verbs;
– 'existance' (entity) verbs are the verbs denoting self-evident and directly perceived existence of a referent;
– 'event' verbs are the verbs denoting active actions, activity, states of activity.

The third class of verbs is a widely versatile system, represented in the following sets: a) verbs denoting mental and emotional activity, as well as the activity of thought and spirit; b) verbs referring to actions related to indivisible spiritual and physical sphere: naming actions - behaviors and contacts, information, as well as actions denoting work, various physical actions and movements. Each of these sets combines cross-branched sections. This class also includes verbs denoting inactive procedural states – physical and physiological. In Tables 2 and 3 we present the classes of verbs which co-occur in all the grade-level texts and, thus, form high-frequency bigrams with the word 'chelovek' (a man) in the Corpus.

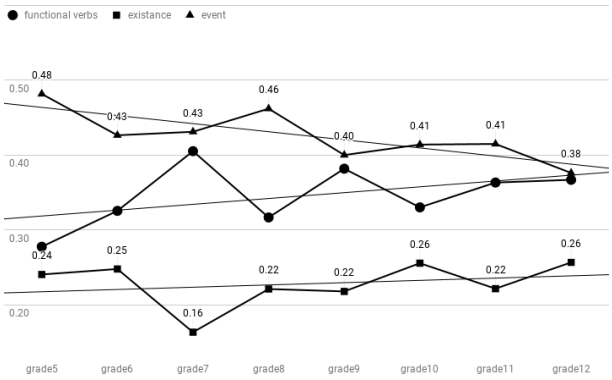

Fig. 1. Three semantic classes of verbs in Textbooks of Gradelevels 5–12

Table 2. Distribution of three semantic classes of verbs in Grade levels 5–12

	Grade							
Verb type	5	6	7	8	9	10	11	12
functional	0.28	0.33	0.41	0.32	0.38	0.33	0.36	0.37
existance	0.24	0.25	0.16	0.22	0.22	0.26	0.22	0.26
event	0.48	0.43	0.43	0.46	0.40	0.41	0.41	0.38

As we see, the core, i.e. words that most frequently collocate with 'chelovek (a man) is made by the semantic classes of functional verbs (Russian byt' (to be), stat' (become)), modal verbs (moch' (can, be able to)), verbs of possession (imet' (to have, possess). The verbs of other semantic classes are less frequent,

Table 3. Diversity of the three semantic classes of verbs in Grade levels 5–12

	Grade							
Verb type	5	6	7	8	9	10	11	12
functional	0.29	0.30	0.33	0.27	0.33	0.26	0.29	0.39
existance	0.18	0.18	0.14	0.13	0.19	0.20	0.13	0.19
event	0.53	0.52	0.53	0.59	0.48	0.53	0.58	0.42

though there are two more verbs characterized with above average frequency, i.e. Russian zhit'(to live) and stremit'sya (to seek). Semantic classes of verbs that collocate with the word 'chelovek' (a man) in texts of Grade 5 textbook are presented below. The most lexically diversified in texts of grade 5 is the class of verbs denoting different types of actions. They are represented by the core vocabulary of the Russian language, the words which posses a high frequency in the National language: live, make, develop, sleep, produce, try, turn, socialize, communicate, answer, etc.

The lexico-grammatical constructions or bigrams of verbs co-occurring with the word 'chelovek' (a man) functioning as a subject in texts of the 11th grade demonstrated a wider range of the verbs used. The twenty highest ranking verbs are the following: can (25), be (15), be (13), have (12), live (5), become (5), strive (4), understand (4), define (4), speak (3), engage in (3), acquire (3), manifest (3), follow (3), anticipate (3), call (3), see (3), create (3), join (3). The contrastive analysis of the bigrams retrieved from the subcorpora of texts of different classes (5 – 11) demonstrate that the variety of the semantic groups of the verbs are the same but the number of the verbs with which the word 'chelovek' (a man) collocates in each group is much higher thus the groups are 'densely populated'. E.g. the groups of verbs denoting existence and status increase dramatically acquiring a wider range of verbs. As it is seen in Fig. 4.1 the number of functional verbs increase over the grade level line from 0.28 in Grade 5 to 0.37 in Grade 12 (which in the Graph marks textbooks of the 11th Grade of the Advanced level).

5 Conclusion

In this paper we report a corpus study aimed at testing the hypothesis that bigram distributional information could be a function of text complexity. In a 623782-word corpus of Russian textbooks on Social studies for middle and high school we explore how the number of bigrams of the core vocabulary correlate with reading levels of texts within the grade range 5 – 11. The applied methods and techniques are exemplified with the bigrams (Noun + Verb) of the most frequent content noun in the corpus, i.e. 'chelovek' (a man). As it is unanimously accepted that (a) texts with low frequency words are more difficult to read and (b) frequency of separate words and bigram has proven high discriminative power among other readability metrics in many languages, in this study we have put our focus on two features providing rich text representation and readability prediction:

– the number of bigrams ‘chelovek’ (a man) + VERB and
– and the semantic range of the verbs in the bigrams ‘chelovek’ (a man) + VERB.

The findings reveal that the distributional patterns of the identified core bigrams construct particular semantic classes which tend to increase from grade to grade. The research results may contribute in the following areas of professional knowledge:

1. Textbook writers on Social sciences may be supplied with a better understanding of the prevalence and type of verbs to be used to generate texts of a certain reading profile.
2. Researchers may be better able to identify type markers and rank reading texts.
3. Identifying the ratio of functional, event (action) and entity (existance) verbs as a marker of text type, genre and complexity can be beneficial for the development of better reading formulas.

The findings are particularly relevant for text complexity theory as they are consistent with the previous results of corpus investigations on correlation of text complexity and a number of text features. The results of the research may also have major implications for natural language processing in text complexity research. The issue which we decided not to address in the present work is granularity of verb classes. It is obvious that the ‘appropriate’ level of class and subclass granularity may vary from one research to another. In the present study we provided a general purpose classification suitable for various purposes, and in the future we intend to refine and organize semantic classes of verbs into taxonomies of higher degrees of granularity.

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 up to subsection 3.2 in the paper) was created without supporting by the Russian Science Foundation.

References

1. L. G. Babenko. *Explanatory Ideographical Dictionary of Russian Verbs.* Moscow, Ast-Press, 1999.
2. A Bhakkad, S.C. Dharamadhikari, and P. Kulkarni. Efficient approach to find bigram frequency in text document using e-vsm. *International Journal of Computer Applications*, 68(19), 2013.

3. E. Dale and J. S. Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54, 1948.
4. O.V. Dereza, D.A. Kayutenko, and A.S. Fenogenova. Automatic morphological analysis for Russian: A comparative study. 2016.
5. W. B. Elley. The assessment of readability by noun frequency counts. *Reading research quarterly*, pages 411–427, 1969.
6. M. Flor, B.B. Klebanov, and K. M. Sheehan. Lexical tightness and text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 29–38, 2013.
7. E Fry. Readability: Insights, sidelights, and hindsight. *JV Hoffman & YM Goodman (Red.), Changing literacies for changing times: an historical perspective on the future of reading research, public policy, and classroom practices*, pages 174–185, 2009.
8. E. J. Gibson, A. Pick, H. Osser, and M. Hammond. The role of grapheme-phoneme correspondence in the perception of words. *The American Journal of Psychology*, 75(4):554–570, 1962.
9. V.V. Ivanov, M.I. Solnyshkina, and V.D. Solovyev. Efficiency of text readability features in Russian academic texts. In *Computational Linguistics and Intellectual Technologies*, volume 17, pages 277–287, 2018.
10. G. R. Klare et al. Measurement of readability. 1963.
11. E. Kuzmenko. Morphological analysis for Russian: integration and comparison of taggers. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 162–171. Springer, 2016.
12. B. Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.
13. J. M. Mason. The roles of orthographic, phonological, and word frequency variables on word-nonword decisions. *American Educational Research Journal*, 13(3):199–206, 1976.
14. Philip M McCarthy and Scott Jarvis. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488, 2007.
15. P.D. Pearson and A. Studt. Effects of word frequency and contextual richness on children's word identification abilities. *Journal of Educational Psychology*, 67(1):89, 1975.
16. N.Y. Shvedova. *Russian Semantic Dictionary*, volume 4. Moscow, Azbukovnik, 2007.
17. M. Solnyshkina, E. Harkova, and A. Kiselnikov. Comparative coh-metrix analysis of reading comprehension texts: Unified (Russian) state exam in English vs Cambridge first certificate in English. *English Language Teaching*, 7(12):65, 2014.
18. V. Solovyev, V. Ivanov, and M. Solnyshkina. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent & Fuzzy Systems*, 34(5):3049–3058, 2018.
19. G. Spache. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413, 1953.
20. J. Trier. *Der deutsche Wortschatz im Sinnbezirk des Verstandes: von den Anfängen bis zum Beginn des 13. Jahrhunderts*, volume 31. C. Winter, 1931.
21. A. Wierzbicka. Lingua mentalis: the semantics of natural language. 1980.
22. M. Xia, E. Kochmar, and T. Briscoe. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, 2016.