Quantitative analysis of lexical complexity in contemporary Russian novels

Yu S Maslennikova and A V Abramov
Kazan Federal University, Kremlevskaya str. 18, Kazan 420018, Russia
E-mail: JSMaslennikova@kpfu.ru

Abstract. This research presents a set of methods for measuring the lexical complexity against the culture at large. We applied a range of statistical measures upon a set of texts. For the analysis, we selected 50 famous Russian novels of the 20th century based on their historical and generic variety. The main approach consists of the comparison of the relative frequency of a work’s words against the Google Books dataset. This dataset represents a remarkable resource, with the Russian Google Books corpus in the period 1800 - 2012, containing approximately 4.7 billion 1-grams. Relative frequencies distribution for a novels words were compared with frequencies of Google Books corpus of different years using Jensen-Shannon divergence, Kullback-Leibler divergence and other information measures. Also, so-called a Flesch reading ease scale was calculated, such measure was used in similar research of modern English fiction. It was shown that the lexical complexity of individual texts should be measured against the culture at large. It was found that a writer who commonly seems to be difficult, verbose or notional, can in fact use language that is more ‘common’ (relative to the culture at large) than any other texts.

1. Introduction
The study of literary complexity has a long history. But especially in the modernist period, the complexity has come to be stronger identified with literature, and literature became synonymous with difficulty. At this historical moment, the first strictly formalized approaches to the quantitative study of language arise. The most commonly used of such quantitative measures are the Flesch-Kincaid readability index and Flesch reading ease scale (FRES) that were designed to indicate how difficult a passage in English is to understand. These measures use sentence length and syllable count to estimate the position of the text on a complexity scale [1]. But analyzing of modernist literature, often the semantics and abstractness of the text can be the main factors of complexity. The aim of the most research included an attempt to identify such lexical attributes as complexity, readability, or difficulty [2]. But literature complexity can be defined as an intersection of these terms because of the problem is that complexity itself is like a moving target: what at first feels forbiddingly new can come to seem familiar to the reader of tomorrow [3]. Authors of the research [3] present a numerical method for measuring the complexity of English modern fiction texts against the broader culture at large using the Google Books Fiction dataset. This corpus has the advantage of providing a huge volume of text from which to estimate word frequency or rarity in the period that a text was written [4]. In this research we present a similar approach for measuring the complexity of contemporary Russian novels (late 19th century to the early 21st century). Relative frequencies distribution for a
novel’s words were compared with frequencies of the Russian Google Books corpus of different years using a set of information measures.

2. Methods
Due to the fact that in Russian language, a sentence length is commonly less than in English (because of the less use of service words such as articles or auxiliary verbs), and words are on the average longer, Flesch reading ease scale (FRES) should be adapted for example by fitting the coefficients by comparing the indices obtained for the original English texts and their translations. In this research, the following adaptation of FRES scale was used [5]:

\[ \text{FRES} = 206.835 - 1.3 \cdot \text{ASL} - 60.1 \cdot \text{ASW}, \]

where ASL an average sentence length, ASW an average number of syllables per word. Values of the range 0-30 correspond to the reading level “very difficult to read”, 30-50 is “Difficult to read” etc.

To estimate the lexical complexity of different Russian novels against the culture at large, we set out a set of complexity measures that compare the relative frequencies of a given novels words, as measured against the Russian Google Book dataset. This dataset represents a remarkable resource, with the Russian Google Books corpus in the period 1800 - 2012, containing approximately 4.7 billion 1-grams [4]. First of all, we defined a measure that characterizes an effective volume of the given texts vocabulary using the Shannon entropy:

\[ h_{x/x} = - \sum_i p_x^i \log_2 p_x^i, \]

where \( p_x^i \) is relative frequency of a word \( x_i \) in a text. Generally, the entropy quantifies the expected value of the information contained in a message and it is measured in bits. Lets define the entropy of the Google Books corpus as:

\[ h_{y/y} = - \sum_i p_y^i \log_2 p_y^i. \]

The entropy of the corpus can be estimated by year. The following information measure, as \( h_{x/y} = - \sum_i p_x^i \log_2 p_y^i \) can be defined as an average length of a codeword when the given text is encoded using the Google Books frequency dictionary. It can be shown, that generally \( h_{x/y} \geq h_{x/x} \). Therefore, the difference between the given texts frequency dictionary and the Google Books corpus can be defined as the difference between two entropies:

\[ \text{KLD}_{x/y} = h_{x/y} - h_{x/x}. \]

This divergence is also well-known as the KullbackLeibler divergence. In terms of lexicography, this divergence measures the lexical matter of individual text against the broader culture at large. Instead of the KullbackLeibler divergence, in this research, we will use the JensenShannon divergence which is more widely used in natural language processing applications. It can be define as:

\[ \text{JSD}_{x/y} = \frac{1}{2} \text{KLD}_{x/m} + \frac{1}{2} \text{KLD}_{y/m}, \]

where \( m = \frac{1}{2} (p^x + p^y) \).

Additionally, we will compare novels using the following relative information score:

\[ \Delta H_{x/y} = h_{x/y} - h_{y/y}. \]

The negative value of this score (\( \Delta H < 0 \)) indicates the prevalence of wide-commonly used words in a text and the positive one (\( \Delta H > 0 \)) - the prevalence of uncommon words. This score could show writers who employ language that is more ‘common’ (relative to the culture at large) than any other texts.

3. Results
For the analysis, we have selected Russian literature texts that are diverse in terms of historical range (late 19th century to the early 21st century), a literary genre (comedy, realism, military prose), and authors styles. The full list of novels with years of the first publication is presented in Table 1.

Fig. 1a shows Shannon’s entropy dynamics of the Russian Google Book dataset for the 20th century. It is shown that the entropy has increased by 1 bit approximately, in other words, the effective volume of the common Russian vocabulary has doubled, for example, by an extension of new concepts and a birth of new word forms. On Fig. 1b Shannon’s entropy for all selected Russian novels is presented by year. We don’t see the correlation between increasing of author’s entropy by year, but we can see the significant difference in entropy values (3 bits) for different authors. The minimal entropy value (8.1 bits) corresponds to the novel “An evening with
Claire” (G. Gazdanov, 1929), the maximal value (12.1 bits) - to the novel “The Inscription” (A. Prokhanov, 2006).

Table 1. The list of 50 contemporary Russian novels

| ID | Author | Title                        | ID | Author | Title                        |
|----|--------|------------------------------|----|--------|------------------------------|
| 1  | L. Tolstoy | Hadji Murat(1904)    | 26 | V. Shalamov | Kolyma Tales(1962) |
| 2  | A. Chekhov | The Cherry Orchard(1903)   | 27 | A. Solzhenitsyn | One Day in the Life of Ivan Denisovich(1959) |
| 3  | I. Bunin  | Dark Alleys(1945)          | 28 | K. Vorobyov | Here We Are, My Lord!(1946) |
| 4  | D. Merezhkovsky | Little Teresa(1941) | 29 | V. Bolotin | In the August of ‘44(1973) |
| 5  | A. Kuprin | The Duel(1965)             | 30 | V. Astafiev | The Cured and the Sinned(1993) |
| 6  | M. Gorky  | The Life of Klim Samgin(1936)| 31 | Yu. Bondarev | The Hot Snow(1969) |
| 7  | F. Solzhen | The Petty Demon(1902)      | 32 | Ye. Novov | Ulyanovskie Holotniki(1977) |
| 8  | A. Bely   | Petersburg(1913)           | 33 | V. Shukshin | To the Third Roseters(1975) |
| 9  | V. Rozanov | Apocalypse of Our Time(1917)| 34 | V. Belov | Business as Usual(1966) |
| 10 | Y. Zamyatin | We(1920)               | 35 | V. Rasputin | Farewell to Mataora(1976) |
| 11 | N. Ostrovsky | How the Steel Was Tempered(1934)| 36 | G. Klimov | Prince of this World(1958) |
| 12 | A. Platonov | Chevengur(1928)         | 37 | L. Tolstoy | The Kreutzer Sonata(1889) |
| 13 | L. Leonov | Thief(1927)               | 38 | V. Makshin | Seven Days for an Eternity(1971) |
| 14 | M. Sholokhov | And Quiet Flows the Don(1932)| 39 | V. Erofeev | Moscow - Potshuki(1969) |
| 15 | M. Bulgakov | The Master and Margarita(1937)| 40 | Yu. Kasakov | Northern Diary(1972) |
| 16 | A. Grin   | She Who Runs on the Waves(1928)| 41 | Yu. Mamleev | Shatuny(1968) |
| 17 | A. Fadeev | Raagrom(1926)             | 42 | A. Vampilov | Duck Hunt(1970) |
| 18 | B. Pasternak | Doctor Zhivago(1955)    | 43 | A. Bittov | Pushkin House(1971) |
| 19 | A. Tolstoy | The Road to Calvary(1917)  | 44 | A. Prokhanov | The Inscription |
| 20 | V. Nabokov | The Gift(1958)            | 45 | V. Makarin | Underground or a Hero of Our Time(2006) |
| 21 | V. Kaverin | Two Captains(1944)        | 46 | A. Kim | Father-forest(1989) |
| 22 | V. Nekrasov | Front-line Stalingrad(1945)| 47 | A. Tolstoy | Peter the First(1934) |
| 23 | A. Tvardovsky | Yavlits Tyorkin(1945)    | 48 | E. Lemonov | We Had a Great Era(1989) |
| 24 | G. Gazdanov | An Evening with Claire(1929)| 49 | S. Sokolov | School for Fools(1973) |
| 25 | I. Ilf and Ye. Petrov | The Twelve Chairs(1927) | 50 | Yu. Poljakov | Goat in Milk(1995) |

FRES scale on Fig. 2 shows several trends. Despite the common understanding of contemporary literature as being more expansive than what comes before it is noticeable that most of selected novels have the reading level “difficult to read” or “very difficult to read”. But the distribution of FRES scale shows that this measure has a small correlation with the actual lexical complexity. That’s why on Fig. 3, 4 we present the mean relative information score $\Delta H_{x/y}$ and the mean Jensen-Shannon divergence correspondingly. The calculations only included words that were found both in the novel and in the corpus from the period. The two extremes of Fig. 3 (“Two captains” and “An evening with Claire”) demonstrate those texts that feature, respectively, the least and the most common lexis, relative to the corpus.

Figure 1. (a) Shannon’s entropy dynamics of the Russian Google Book dataset in the 20th century; (b) Shannon’s entropy by year for 50 selected Russian novels
4. Conclusion
In this research, we tried to show that any real attempts to understand quantitative literature complexity need measure against the culture at large. It was shown that a writer who commonly seems to be difficult, verbose or notional, can in fact use language that is more ‘common’ (relative to the culture at large) than any other texts. The answer about lexical complexity never is the only one because of what at first feels forbiddingly new can come to seem familiar to the reader of tomorrow. And we will never quantify such cultural variability. But we presented the approach that utilizes permanently growing indicators of the culture at large like Google Books. Also, it can be seen that quantitative analysis of contemporary Russian literature is only on the initial stage.

This work was supported by the Russian Foundation for Basic Research, Grant no. 17-29-09163. The research of the first author was supported by the Russian Government Program of Competitive Growth of Kazan Federal University.

References
[1] Rudolf Flesch 1948 A new readability yardstick, Journal of Applied Psychology, 32
[2] Pitler, E and Nenkova, A 2008 Revisiting readability: a unified framework for predicting text quality (In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii: Association for Computational Linguistics)
[3] Ewan Jones, Paul Nulty 2019 Quantitative measures of lexical complexity in modern prose fiction, Digital Scholarship in the Humanities, https://doi.org/10.1093/lhc/fq20
[4] Michel J, Shen Y, Aiden A, Veres A, Gray M, The Google Books Team, Pickett J, Hoiberg D., Clancy D., Norvig P, Orwajg J, Pinker S, Nowak M, Aiden E 2011 Quantitative Analysis of Culture Using Millions of Digitized Books Science 331 no. 6014 , p. 176-182
[5] Oborneva, I V 2006 Automated estimation of complexity of educational texts on the basis of statistical parameters Pedagogy Cand. Diss. Moscow (In Russian)
[6] Hinrich Schutze, Christopher D Manning 1999 Foundations of Statistical Natural Language Processing (Cambridge, Mass: MIT Press)