Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

MODELING LEMMA FREQUENCY BANDS FOR LEXICAL COMPLEXITY ASSESSMENT OF RUSSIAN TEXTS[1]

Blinova O. V. (o.blinova@spbu.ru; ovblinova@hse.ru)
Saint Petersburg State University;
HSE University, Saint Petersburg, Russia

Tarasov N. A. (tarasovn2468@yandex.ru),
Modina V. V. (st072157@student.spbu.ru),
Blekanov I. S. (i.blekanov@spbu.ru)
Saint Petersburg State University, Saint Petersburg, Russia

The paper is devoted to the problem of modeling general-language frequency using data of large Russian corpora. Our goal is to develop a methodology for forming a consolidated frequency list which in the future can be used for assessing lexical complexity of Russian texts.

We compared 4 frequency lists developed from 4 corpora (Russian National Corpus, ruTenTen11, Araneum Russicum III Maximum, Taiga). Firstly, we applied rank correlation analysis. Secondly, we used the measures "coverage" and "enrichment". Thirdly, we applied the measure "sum of minimal frequencies". We found that there are significant differences between the compared frequency lists both in ranking and in relative frequencies. The application of the "coverage" measure showed that frequency lists are by no means substitutable. Therefore, none of the corpora in question can be excluded when compiling a consolidated frequency list.

For a more detailed comparison of frequency lists for different frequency bands, the ranked frequency list, based on RNC data, was divided into 4 equal parts. Then 4 random samples (containing 20 lemmas from each quartile) were formed.

Due to the wide range of values, accepted by ipm measure, relative frequency values are difficult to interpret. In addition, there are no reliable thresholds separating high-frequency, mid-frequency, and low-frequency lemmas. Meanwhile, to assess the lexical complexity of texts, it is useful to have a convenient way of distributing lemmas with certain frequencies

[1] The presented research was supported by the Russian Science Foundation, project #19-18-00525 "Understanding official Russian: the legal and linguistic issues".

over the bands of the frequency list. Therefore, we decided to assign lemmas “Zipf-values”, which made the frequency data interpretable because the range of measure values is small.

The result of our work will be a publicly accessible reference resource called “Frequentator”, which will allow to obtain interpretable information about the frequency of Russian words.

Key words: Russian, linguistic corpora, lemma frequency lists, general-language frequency, frequency bands, low-frequency words, lexical complexity

DOI: 10.28995/2075-7182-2020-19-76-92

МОДЕЛИРОВАНИЕ ЗОН ЧАСТОТНОГО СЛОВАРЯ ДЛЯ ОЦЕНКИ ЛЕКСИЧЕСКОЙ СЛОЖНОСТИ РУССКИХ ТЕКСТОВ

Блинова О. В. (o.blinova@spbu.ru; ovblinova@hse.ru)
Санкт-Петербургский государственный университет;
НИУ «Высшая школа экономики», Санкт-Петербург, Россия

Тарасов Н. А. (tarasovn2468@yandex.ru),
Модина В. В. (st072157@student.spbu.ru),
Блеканов И. С. (i.blekanov@spbu.ru)
Санкт-Петербургский государственный
университет, Санкт-Петербург, Россия

Introduction

The study is aimed at the problem of forming a consolidated lemma frequency list based on the frequency lists of large Russian corpora. Such a list can be used to assess the lexical complexity of Russian texts (for example, it will be possible to estimate the number of low-frequency, i.e. unfamiliar, words of the text and use these values in readability formulas). Such a list should contain interpretable frequency values that will allow us to divide the frequency list into bands and distinguish between high-frequency, mid-frequency and low-frequency lemmas.

Section 1 discusses readability formulas that take into account the number of long words or (un)familiar words; it is concluded that the application of the familiarity criterion is difficult to operationalize without reference to word frequency data. Section 2 shows that features including word frequency information successfully predict text complexity. Section 3 discusses general-language frequency and the

problem of accounting for the reader's actual language experience. **Section 4** briefly discusses approaches to identifying frequency bands. **Section 5** gives a description of four Russian corpora, whose frequency lists are involved in the comparison. Section 6 describes the methods for comparing frequency lists; **section 7** gives the results of applying the selected methods. The results indicate that there are significant differences between the compared frequency lists both in the ranks of the lemmas and in their relative frequencies, and that the frequency lists are not substitutable. **Section 8** justifies the use of the frequency measure "Zipf-value" which has a small range of values.

1. Long or unfamiliar words and texts complexity

There is a fairly long tradition of applying readability assessment methods to texts in Russian; for a review see [Reynolds 2016]. In particular, readability metrics are used, that is, formulas where variables include the number of complex words. Complex words can be understood either as long (multicharacter or multisyllabic) units, or as unfamiliar units.

Although, as K. Collins-Thompson pointed out, "the word lists used in vocabulary-based readability measures like Dale-Chall may be thought of as a simplified language model" [Collins-Thompson 2014], see also [Crossley et al. 2019], the use of such formulas is a common method for assessing the document complexity. Presently it is used in combination with other, more sophisticated methods, for more details see, for example [Benjamin 2012]. More precisely, the number of complex (long, unfamiliar/rare/low-frequency) words of the text or the average length of words in letters or syllables is used in various text classification models as one of many features, see, e.g., [Schwarm, Ostendorf 2005].[2] It is clear that, with the exception of some special cases,[3] the application of the familiarity criterion is difficult or impossible to operationalize without using word frequency information.[4]

2. Word frequency as a parameter for text complexity assessing

According to [Leroy, Kauchak 2014], the word frequency is closely related to both the actual word complexity (measured by how well readers can choose the correct definition of the word) and the difficulty to read.

---

[2] Recent studies show that "sentence and word length measures likely do not tap directly into linguistic components related to readability" [Crossley et al. 2019]. However, it is clear that the various parameters for assessing lexical complexity are not independent of each other, in particular, according to Zipf's law of abbreviation, the length of a word correlates with its frequency, see, for example, [Bentz, Ferrer-i-Cancho 2016].

[3] These are cases with "lexical minimums" or with the results of painstaking surveys aimed at identifying familiar words.

[4] For example, in [Batinić et al. 2016] and in "LeStCor: Levelled Study Corpus of Russian" the words included in the list of 5000 most frequent Russian words compiled by S. A. Sharoff [Sharoff, electronic resource], see also [Sharoff et al. 2013], are treated as familiar.

The studies of Russian text complexity for native speakers or second language learners also show that lexical features, including information on word frequency and/or inclusion in vocabulary lists for each CEFR level ("lexical minimums"), successfully predict complexity. For instance, according to [Laposhina 2017], it is precisely these features that showed the highest correlation with complexity. In [Ivanov et al. 2018] metrics based on lexical features (including word frequency, average frequency of nouns, etc.) are evaluated as reliable, see also [Sharoff et al. 2008], [Solovyev et al. 2018].

Frequency information can be applied in various ways. The average absolute word frequency or mean log frequency [Collins-Thompson, Callan, 2005], the total frequency of content words [Inavov et al. 2018] etc. can be used as measures of lexical complexity. In addition, when assessing text complexity, one can take into account the number of words that are not included in the lists of (high)frequency words, for more details on more sophisticated models, see [Chen, Meurers 2016].

Lemma frequency can be estimated using frequency dictionaries or representative corpora. In this paper, we focus on the problem of the general-language frequency modeling based on data from large Russian corpora.

3. In search of general-language frequency

According to K. Collins-Thompson, "a widely-used feature of lexical difficulty for a word is thus the relative frequency of that word **in everyday usage**,[5] as measured by its relative frequency in a **large representative corpus**, or its presence/absence in a **reference word list**" [Collins-Thompson 2014]. To assess the general-language frequency of words, one should use some "general-language corpus", see the studies on designing and balancing corpora and corpora representativeness, e.g., [Atkins et al. 1992]. As stated in [Biber 1993: 247], a representative corpus "might contain roughly 90% conversation".[6]

In [Chen, Meurers 2016] this problem of accounting for the actual competence of a native speaker is also discussed, cf.: "the frequency lists adopted by these studies were mostly drawn from written corpora. Spoken language was rarely taken into consideration when frequency lists were being composed. This runs the risk of the frequency values not being a faithful representation of the reader's actual language experience, hence being suboptimal for predicting the ease of perception and retrieval". Accordingly, when modeling the general-language frequency for Russian it would be reasonable to give greater weight to the frequency values, obtained from a spoken corpus (e.g., Corpus of Spoken Russian in the Russian National Corpus).

---

[5] See also citation from [Slioussar 2005]: "Many psycholinguists who use data on the frequency of certain words or forms are often subjected to harsh criticism. After all, such data is most often taken from frequency dictionaries, based exclusively on written texts, not oral ones. Even to a layman it is intuitively clear that the frequency of words and their forms in colloquial speech should correlate with the frequency presented in the mental lexicon".

[6] As far as we know, balanced corpora organized according to the indicated principle have not been created yet.

4. Methods for modeling general-language frequency and frequency bands

The word frequency effect studies demonstrate that high-frequency words are usually perceived and produced more efficiently and faster than low-frequency ones, see, for example, [Brysbaert et al. 2018].

Meanwhile, if we use classical techniques for text complexity prediction using frequency information, averaging over all frequency values, then the contribution of low-frequency words becomes minimal [Chen, Meurers 2016]. Therefore, we are faced with the task of identifying frequency bands that explicitly show high-frequency, low-frequency, and mid-frequency units.

Various thresholds values (for the frequencies or ranks) are used to separate the bands.[7] The conventional threshold value for low-frequency words in a 100 million word corpus is 5 ipm (items per million) [Lyashevskaya 2016: 236]. Different threshold values are also used for ranks. High-frequency units are the words with a rank up to 2,000 [Schmitt 2010, 69]; mid-frequency units are words with ranks from 2,000 to 8,000–9,000 [Schmitt 2010: 70]. Rare units in the New Frequency Dictionary of Russian are the lemmas with a rank of 10,000 and more [Lyashevskaya 2016: 229]. The entire frequency list can be divided into quartiles (for example, in [Zhao, Jurafsky 2009] words from the lower quartile of the ranked frequency list are considered as low-frequency ones); percentiles can also be used for this purpose, see [Bell et al. 2009].

In this paper we compare 4 frequency lists based on four Russian corpora. These corpora are of different size and composition. Our goal is to develop a methodology for creating a consolidated lemma frequency list based on the frequency lists of large Russian corpora.

5. Frequency data sources

This paper compares frequency lists derived from three large web corpora: ruTenTen11 [ruTenTen11, electronic resource], [Kilgariff et al. 2014], Araneum Russicum III Maximum [Araneum Russicum, electronic resource], [Benko 2014], Taiga [Taiga, electronic resource], [Shavrina, Shapovalova 2017] and the New Frequency Dictionary of Russian (NFDR), based on data from Russian National Corpus [RNC, electronic resource], [Lyashevskaya, Sharoff 2009].

Frequency lists were obtained from the corpora sites or from corpora creators.[8] In the current version of the Sketch Engine, it was possible to download word lists no longer than 1,000 lines. Therefore, to obtain the most complete frequency list from ruTenTen11, frequency lists of lemmas starting with possible two-letter combinations (*аб*, *ав*, *аг* etc.) were downloaded. The list of possible combinations is obtained using NFDR. For single-letter lemmas, a separate search was performed.

---

[7] It should also be added that low-frequency words are included into the dictionaries of rare, forgotten, uncommon and obsolete words, see, for example, [Somov 1996], [Glinkina 1998], [Ilinskaja 1989], [Rogozhnikova 1997], [Korpusnoj slovar' redkih slov, electronic resource].

[8] The authors of this paper would like to thank Tatyana Shavrina for the opportunity to use the frequency list of the Taiga corpus.

Table 1. Frequency data sources

Corpus	Composition	Size	Analyser	Number of lemmas in the frequency list
RNC (NFDR)	genre-balanced RNC subcorpus	91,982,416 graphic words	Mystem	52,138 lemmas with relative frequency ≥ 0.4 ipm (37 occurrences)
ruTenTen11	Internet: news and commercial sites, blogs, social media	near 18 billion tokens (14,553,856,113 text forms)	Treetagger	457,473 lemmas with absolute frequency ≥ 5
Araneum Russicum III Maximum	Internet: news and commercial sites, blogs, social media	15,961,200,372 words	Treetagger	8,893,947 units with absolute frequency ≥ 5
Taiga	Internet: 77% of literary texts (the articles from 33 literary magazines), 19% of naive poetry, 2% of news (from 4 popular news sites), 2% of other texts (popular science, texts of social networks, etc.)	near 5 billion words	UDPipe	2,988,610 lemmas with absolute frequency ≥ 1

6. Methods for frequency list comparison

There are a number of ways to compare frequency lists and methods for measuring the distance between them. In particular, there are measures based on geometrical notions (Euclidean distance, Manhattan distance, Cosine distance, etc.), measures based on well-known statistical tests and procedures (Chi-Square-based measures, Log-Likelihood, Spearman's ρ, etc.), information theoretic measure "perplexity", measure of distance by keywords (Simple Maths) and others, see [Kilgarriff, Rose 1998], [Piperski 2018], [Gomaa, Fahmy 2013] and many others. We chose three measures that allowed us to look at the differences between frequency lists from different points of view (comparing ranks of lemmas, the values of relative frequencies or estimating overlap between the lists).[9]

Firstly, we applied the **rank correlation analysis**, calculating the values of the Spearman and Kendall rank correlation coefficients for pairs of frequency lists. The lists were compared by intersecting lemmas, which equalized their length.

[9] According to [Piperski 2018], the preferred frequency-based measure of corpus distance is Euclidean distance, as this measure is the most robust to corpus size. At the same time, to achieve the objectives of this article, it is sufficient to apply the three measures we have chosen. In addition, some measures (Spearman's ρ, Chi-Square) are commonly used, that is, their application will allow ones comparing our results with the results obtained earlier, see [Khokhlova 2016].

Secondly, we applied two **measures of overlap** ("**Coverage**" and "**Enrichment**"), considered in [Baroni et al. 2009]. The Coverage measure is calculated by the formula:

$$Coverage(X,Y) = \frac{(N1 \cap N2)}{N1}, \quad (1)$$

where X, Y are the corpora, $N1$ is the number of lemmas with an absolute frequency greater than or equal to a given cutoff value in the corpus X, $N2$ is the number of lemmas with an absolute frequency greater than or equal to a given cutoff value in the corpus Y.

The Enrichment measure is calculated by the formula:

$$Enrichment(Y,X) = \frac{M2}{M1}, \quad (2)$$

where $M2$ is the number of lemmas with a frequency above the threshold in the corpus Y and below the threshold in the corpus X, $M1$ is the number of lemmas with an absolute frequency below the threshold in corpus X. As a threshold value, we (following [Baroni et al. 2009]) used the absolute frequency of 20 occurrences. This is the so-called "Sinclair threshold". This (apparently arbitrary) threshold was chosen under the influence of J. Sinclair's statement that an experienced lexicographer would need at least 20 occurrences of an unambiguous word to make a description of its behavior, see, for example, [Lüdeling, Kytö 2009: 818].

Thirdly, we applied the measure "**Sum of Minimum Frequencies**" (SMF), proposed by A. Ya. Shaikevich in [Shaikevich 2015], see also [Piperski 2017]. SMF is calculated by the formula:

$$SMF(X,Y) = \frac{\sum_{min}(pX_i, pY_i)}{\sum_{0.5}(pX_i, pY_i)}, \quad (3)$$

where pX_i is the relative frequency of the lemma in the corpus X, pY_i is the relative frequency of the lemma in the corpus Y.

7. Comparison results

The frequency lists under consideration did not undergo any special preprocessing. Table 2 shows the results of applying rank correlation analysis.

Table 2. Spearman's ρ and Kendall's τ values

Spearman's ρ				Kendall's τ			
X/Y	ruTenTen11	Taiga	NFDR	*X/Y*	ruTenTen11	Taiga	NFDR
Araneum	0.033	0.081	0.223	**Araneum**	0.022	0.006	0.157
ruTenTen11		0.071	0.828	**ruTenTen11**		0.048	0.648
Taiga			0.095	**Taiga**			0.065

The rank correlation coefficient ρ takes value > 0.7 only in the pair ruTenTen11—NFDR ($\rho = 0.828$). This can be explained by the fact that these lists are the shortest and do not contain very long low-frequency tails. In pairs of web-corpora, the correlation coefficients values do not exceed 0.3, that is, the differences in ranking across these corpora are significant.

Table 3 shows the comparison results using Coverage and Enrichment measures. Coverage is a measure of the proportion of words for which there is "enough" information in the corpus *X* and "enough" information in the corpus *Y* [Baroni et al. 2009]. In other words, this is "a (very rough) measure of the extent to which *X* is 'substitutable' with *Y*" [Ibid.]. Enrichment allows one to estimate the proportion of words among those words that are attested in the corpus *X*, and for which there is not enough information in the corpus *X*, but enough information in the corpus *Y* [Ibid.].

Table 3. Values of the measures of overlap, threshold = 20[10]

Coverage				Enrichment			
X/Y	Araneum	ruTenTen11	Taiga	*X/Y*	Araneum	ruTenTen11	Taiga
Araneum		53	51.5	**Araneum**		0.9	0.2
ruTenTen11	7.8		23.1	**ruTenTen11**	3.4		1.9
Taiga	4.6	14.1		**Taiga**	13.9	0.2	

When interpreting presented metrics values, it should be taken into account that the measures are able to evaluate the ratio of frequency lists as *X/Y* or as *Y/X*. The Coverage measure has the highest value for the pairs Araneum (*X*)—ruTenTen11 (*Y*) (53) and Araneum (*X*)—Taiga (*Y*) (51.5); the proportion shows that only about half of the words above the cutoff in Araneum are also above the cutoff in ruTenTen11 and Taiga. Thus, the vocabularies of the compared web corpora are significantly different. The Enrichment values allow one to assess the extent to which the frequency lists are capable of enriching each other. The highest value measure is found for the Taiga—Araneum pair (13.9). Thus, if we consider the entire frequency range in question, the use of various web-corpora is not so beneficial.

On the whole, the assessment of the overlap allows us to conclude that the frequency lists are not substitutable, and when compiling a consolidated frequency list of lemmas, all compared frequency lists should be used.

Finally, Table 4 shows the results of comparing all four lists using SMF measure. This measure compares relative frequencies of all intersecting elements (lemmas) in the lists in pairs.

Table 4. Values of SMF measure

X/Y	ruTenTen11	Taiga	NFDR
Araneum	0.056	0.024	0.264
ruTenTen11		0.116	0.756
Taiga			0.197

Particular attention should be paid to the results of the comparison of web corpora with NFDR. The high value we observe in the pair NFDR—ruTenTen11 (SMF = 0.756). We saw earlier that the rank correlation coefficients for this pair also take the largest

[10] We did not include NFDR in the comparison, since this list contains lemmas with a relative frequency of 0.4 ipm or more (that is, an absolute frequency ≥ 37).

value from the observed values. Significantly less similar are NFDR and Araneum (SMF = 0.264), NFDR and Taiga (SMF = 0.197). This can also be explained by the fact that the frequency lists of Araneum and Taiga contain long tails of low-frequency units.

Thus, applying three measures, we found out that there is significant discrepancy across the lists in ranking and in relative frequencies. The use of the Coverage measure showed that frequency lists are by no means substitutable. Therefore, none of the corpora in question can be excluded when compiling a consolidated frequency list.

8. Comparison by frequency bands

For a more detailed comparison of frequency lists by different frequency bands, we decided to proceed as follows. We divided the ranked NFDR frequency list into 4 equal parts, then, using the ranks, we formed 4 random samples (containing 20 lemmas from each quartile). For each lemma of 4 random samples, we assigned the values of relative frequencies according to all the compared lists. The data obtained for the upper and lower quartiles are presented in **Table 7** and **Table 8** below.

We see that even for lemmas from the upper quartile, there are significant differences in the ipm values according to different corpora. So, the range of ipm values for the most frequent lemma in the sample (the noun *центр* 'centre') is 390.80.

It is important that the **overall range of ipm values** is very significant. NFDR contains lemmas with relative frequencies from 35,801.8 (the conjunction *и* 'and') to 0.4 ipm, Taiga includes lemmas with a frequency from 18,710.7 (the preposition *в* 'in, to, into') to 0.0017 ipm. A significant number of lemmas have frequencies <1 ipm. For example, the Taiga frequency list of 2,988,608 lines contains only 28,500 lemmas with a frequency of ≥ 1 ipm (and this is less than 1/100 of the entire list). The observed proportion of rare words is a consequence of the Zipf's law.

Due to the wide range of values, the observable values of relative frequency are **difficult to interpret**. In addition, there are no reliable thresholds separating high-frequency, mid-frequency, and low-frequency words. Meanwhile, it is useful to have a convenient way of assigning lemmas to certain frequency bands.

Therefore, we (following [Chen, Meurers 2016]) decided to use the approach from [Van Heuven et al. 2014], where a new "Zipf-value" measure of frequency is proposed. The value of this measure is calculated by the formula (4).

$$Zipf\text{-}value = \log_{10}(ipm \times 1000), \tag{4}$$

The measure has the following advantages, see [Ibid.].

1. A logarithmic scale is used.[11]
2. The values are easy to interpret. For example, the most frequent word in NFDR *и* 'and' has Zipf-value equal to 7.55 (or, when rounding to an integer, 8). The word with the lowest frequency in NFDR will have a Zipf-value of 2.6 (or 3).

[11] The values of the logarithmic frequencies are used by psycholinguists, see for example, [Winter 2020, 95].

3. The scale allows us to separate mid-frequency words from high-frequency and low-frequency ones.
4. Zipf-values are easy to calculate if we know ipm values.

The discussed approach is not the only one possible. In [Sharoff et al. 2017] another logarithmic measure of the frequency "FClass" is proposed (see the formula (5), where freq(max) is the absolute frequency of most frequent word (MFW) in a particular corpus, $freq(w)$ is the absolute frequency of the word in a particular corpus, for which the measure value is calculated).[12]

$$FClass(w) = \log_2 \frac{freq(max)}{freq(w)}, \quad (5)$$

FClass measure also has a small range of values. For example, the lemma *субпопуляция* 'subpopulation' from the lower quartile of NFDR frequency list will take FClass values equal to 16 and 21 (see **Table 5**).

Table 5. FClass values

	freq (*субпопуляция*)	MFW	*freq* (*max*)	FClass
NFDR	37	*и* 'and'	3,293,765.6	16
Taiga	5	*в* 'in, to, into'	11,076,749	21
Araneum	194	*и* 'and'	563,822,183	21

The upper FClass value can be estimated at $freq(w) = 1$, the range of measure values for the compared corpora is [0;22], or [0;23], or [0;29], see. **Table 6**.

Table 6. Maximum FClass values

	freq (*w*)	freq (*max*)	FClass
NFDR	1	3,293,765.6	22
Taiga	1	11,076,749	23
Araneum	1	563,822,183	29
ruTenTen11	1	503,894,565	29

The range of FClass values is greater than the range of Zipf-value. FClass scale does not look like typical Likert rating scale [Jamieson 2004]. Accordingly, interpreting Zipf-values is a simpler task.

Compared frequency lists, as shown below (see **Fig. 1**), obey exponential law. Therefore, we can use Zipf-value as a frequency measure.

[12] The authors would like to thank the anonymous reviewer for pointing out this measure.

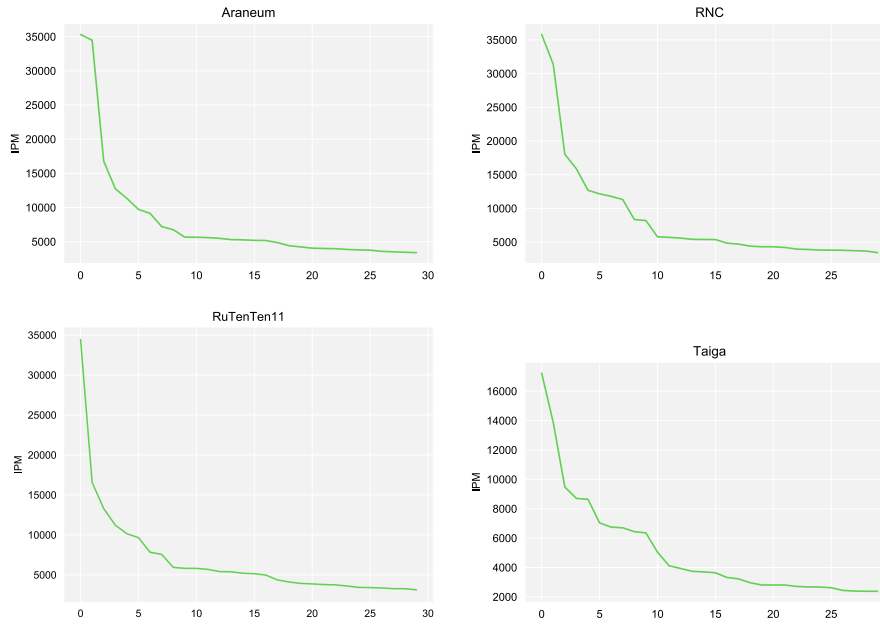

Fig. 1. Frequency distribution

Table 7 and 8 show the Zipf-values calculated for the ipm values in particular frequency lists, as well as mean values, which we will interpret. These values range from 6 (*центр*) to 2 (*субпопуляция*).

Table 7. Lemmas from the upper quartile of the NFDR list

lemma, PoS, translation	ipm				Zipf-value				
	NFDR	Taiga	Araneum	ruTenTen	NFDR	Taiga	Araneum	ruTenTen	mean
центр, N, 'centre'	265.9	187.28	577.41	578.07	5	5	6	6	**6**
участок, N, 'area; district, ward; plot'	144.2	88.32	299.34	273.94	5	5	5	5	**5**
круглый, A, 'round'	71.9	24.96	70.43	73.15	5	4	5	5	**5**
памятник, N, 'monument'	63.8	61.44	81.81	82.62	5	5	5	5	**5**
превратиться, V, 'to transform into'	63.5	0.49	46.47	40.17	5	3	5	5	**4**
чемодан, N, 'suitcase'	42.8	10.24	10.72	11.28	5	4	4	4	**4**
туалет, N, 'toilet, lavatory'	35.1	20.33	31.46	32.12	5	4	4	5	**4**
волшебный, A, 'magic'	28.2	12.92	37.42	31.17	4	4	5	4	**4**
пилот, N, 'pilot'	26.9	14.51	20.36	27.5	4	4	4	4	**4**

lemma, PoS, translation	ipm				Zipf-value				
	NFDR	Taiga	Araneum	ruTenTen	NFDR	Taiga	Araneum	ruTenTen	mean
привлечение, N, 'attraction'	26	14.26	64.41	63.68	4	4	5	5	**5**
ласково, Adv, 'tenderly'	23.6	6.69	4.43	5.7	4	4	4	4	**4**
заказывать, V 'to order'	14.5	15.94	38.5	74.1	4	4	5	5	**4**
взорваться, V, 'to implode'	14.1	1.02	5.43	6.1	4	3	4	4	**4**
вытягивать, V, 'to outstretch; to pull out'	9	9.55	4.66	11.54	4	4	4	4	**4**
Виноградов, N, 'Vinogradov'	7.9	2.87	3.8	5.73	4	3	4	4	**4**
селедка, N, 'herring'	7.3	2.14	2.78	2.11	4	3	3	3	**3**
прибить, V, 'fasten (by nailing)'	7.2	0.09	2.22	0	4	2	3	—	**3**
растворяться, V, 'to dissolve'	7.2	9.04	5.97	6.76	4	4	4	4	**4**
овощной, A, 'vegetable'	6.6	0.65	12.46	12.3	4	3	4	4	**4**
девяностый, Num, 'ninetieth'	6.1	4.47	0.04	4.51	4	4	2	4	**3**

It should be noted that lemmatizers assign different lemmas to the forms of Russian verbs, cf. *превратиться* (Pf)—*превращаться* (Impf), see [Lyashevskaya 2016: 228] about this problem. This is one of the reasons for discrepancies between the frequency lists. The lemma *превратиться* is present in all frequency lists, but in the Taiga list *превратиться* (Pf) has ipm = 0.49, while the lemma *превращаться* (Impf) has ipm = 55.36, which is much closer to the values demonstrated by others corpora. Similar discrepancies in the ipm values are observed for lemmas *взорваться* (*взрываться*) and *прибить* (*прибивать*).

The list of lemmas from the second quartile can be commented on in the same way as the list of lemmas from the first one. In the ruTenTen11 list the lemma *подоспеть* (Pf) 'arrive in time' was not found, but there was the lemma *подоспевать* (Impf). Lemmas from the second quartile (three of which have an average Zipf-value equal to 4, 16 have a Zipf-value equal to 3, 1 (*окрылить* 'inspire') has a Zipf-value equal to 2) for the most part can be considered as mid-frequency ones. The list of lemmas from the third quartile is also quite homogeneous: 15 out of 20 lemmas (75%) have a Zipf-value of 3.

Some low-frequency lemmas from the lower quartile (translation is given in the **Table 8**) cannot be found in two frequency lists of four (*послепожарный*, *тире*), or one frequency list (*несолоно*, *экономразвитие*, *напряг*, *поубавить*, *промельк*, *субпопуляция*). This fact can be explained by lemmatization errors. For instance, representations of the lemma *роздых* in various cases (except for the nominative) are present in the Araneum frequency list.

Accordingly, before the preprocessing of frequency lists for the purpose of forming a consolidated list, it is necessary to decide how to deal with such occurrences as *роздыха*, *роздыху* etc. Apparently, to such occurrences should be assigned normalized forms, and the frequencies of different word forms, related to the same lemma, should be summarized.

Table 8. Lemmas from the lower quartile of the NFDR list

lemma, PoS, translation	ipm				Zipf-value				
	NFDR	Taiga	Araneum	ruTenTen	NFDR	Taiga	Araneum	ruTenTen	mean
тире, N, 'dash'	0.8	0	1.09	0	3	—	3	—	**3**
тявкать, V, 'to yap'	0.7	0.69	0.15	0.28	3	3	2	2	**3**
хроматин, N, 'chromatin'	0.7	0.02	0.05	0.27	3	1	2	2	**2**
линейно, Adv, 'linearly'	0.6	0.24	0.96	1.05	3	2	3	3	**3**
несолоно, Adv, lit. 'unsaltedly'	0.6	0.16	0.1	0	3	2	2	—	**2**
отжимание, N, 'press-up; pressing out'	0.6	0.24	2.2	0.23	3	2	3	2	**3**
пеленг, N, 'bearing'	0.6	0.06	0.15	0.39	3	2	2	3	**2**
денатурация, N, 'denaturating'	0.5	0.01	0.06	0.18	3	1	2	2	**2**
подледный, A, 'subglacial'	0.5	0.34	0.34	0	3	3	3	—	**3**
роздых, N, 'rest'	0.5	0.15	0	0.15	3	2	—	2	**2**
сахарок, N, 'sugar' (diminutive)	0.5	0.11	0.08	0.17	3	2	2	2	**2**
экономразвитие, N, 'economic development'	0.5	0.06	0.13	0	3	2	2	—	**2**
буерак, N, 'ravine'	0.4	0.15	0.01	0.25	3	2	1	2	**2**
втык, N, 'tongue-lashing'	0.4	0.09	0.08	0.1	3	2	2	2	**2**
депонировать, V, 'to deposit'	0.4	0.04	0.05	0.31	3	2	2	2	**2**
напряг, N, 'stress'	0.4	0.96	0.41	0	3	3	3	—	**3**
послепожарный, A, 'post-fire'	0.4	0.03	0	0	3	1	—	—	**2**
поубавить, V, 'to diminish'	0.4	0.08	0.1	0	3	2	2	—	**2**
промелькь, N, 'flash'	0.4	0.3	0.01	0	3	2	1	—	**2**
субпопуляция, N, 'subpopulation'	0.4	0.01	0.01	0	3	1	1	—	**2**

Conclusion

Thus, we compared the frequency lists derived from four Russian corpora. Our aim was not comparison itself, but the development of a methodology for creating a consolidated frequency list and modeling the general-language frequency. It seems that the inclusion of Zipf-value in such a list will make the frequency data interpretable, since the range of measure values is small (the most frequent lemmas will have Zipf-values equal to 7 and 8, the least frequency lemmas will have Zipf-values equal to 1 and 2).

The result of our work[13] will be a publicly accessible reference resource called "Frequentator" which will allow to obtain interpretable information about the frequency of Russian words. To create such a resource, it will be necessary to preprocess

[13] The authors would like to express their sincere gratitude to anonymous reviewers for useful comments regarding the upcoming work on the consolidated frequency list formation.

the frequency lists of web corpora, detect and remove noise; perform lemmatization of occurrences that do not coincide with normalized forms; assign to each lemma a part-of-speech tag; analyze verbs and form a consolidated list. At the end, each lemma will be assigned a weighted frequency value in ipm and Zipf-value.

References

1. *Araneum Russicum III Maximum*, available at: http://ucts.uniba.sk/aranea_about/_russicum.html.
2. *Atkins S., Clear J., Ostler N.* (1992), Corpus Design Criteria, Literary and Linguistic Computing, Vol. 7, № 1, pp.1–16.
3. *Baroni M., Bernardini S., Ferraresi A. & Zanchetta E.* (2009), The WaCky wide web: a collection of very large linguistically processed webcrawled corpora, Language Resources and Evaluation, 43, pp. 209–226.
4. *Batinić D., Birzer S. & Zinsmeister H.* (2016), Creating an extensible, levelled study corpus of Russian, Dipper S., Neubarth F., Zinsmeister H. (eds.), Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), pp. 38–43. (Bochumer Linguistische Arbeitsberichte 16).
5. *Bell A., Brenier J. M., Gregory M.,Girand C., Jurafsky D.* (2009), Predictability effects on durations of content and function words in conversational English, Journal of Memory and Language, 60, pp. 92–111.
6. *Benjamin R. G.* (2012), Reconstructing readability: recent developments and recommendations in the analysis of text difficulty, Educational Psychology Review, 24(1), pp. 63–88.
7. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. P. Sojka, A. Horák, I. opeček and K. Pala (Eds.). Text, Speech and Dialogue. 17th International Conference, TSD 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264.
8. *Bentz C. & Ferrer-i-Cancho R.* (2016), Zipf's law of abbreviation as a language universal, Bentz C., Jager G. & Yanovich I. (eds.) Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. University of Tubingen, online publication system, available at: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.
9. *Biber D.* (1993), Representativeness in Corpus Design, Literary and Linguistic Computing, Vol. 8, No. 4, pp. 243–257.
10. *Brysbaert M., Mandera P., Keuleers E.* (2018), The Word Frequency Effect in Word Processing: An Updated Review, Current Directions in Psychological Science, Vol. 27, Iss. 1, pp. 45–50.
11. *Chen X., Meurers W. D.* (2016), Characterizing Text Difficulty with Word Frequencies, Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84–94.
12. *Collins-Thompson K.* (2014), Computational assessment of text readability: a survey of current and future research, François Th. and D. Bernhard (eds.), Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics, 165:2, pp. 97–135.

13. *Collins-Thompson K., Callan J.* (2005), Predicting Reading Difficulty with Statistical Language Models, Journal of the American Society for Information Science and Technology, 56(13), pp. 1448–1462.
14. *Corpus dictionary of rare words* [Korpusnoj slovar' redkih slov], available at: http://ruscorpora.ru/new/rarewords.html.
15. *Crossley S. A., Skalicky S., Dascalu M.* (2019), Moving beyond classic readability formulas: new methods and new models, Journal of Research in Reading, 42, 3–4, pp. 541–561.
16. *Glinkina L. A.* (1998), Illustrated dictionary of forgotten and difficult words of Russian literature of XVIII–XIX centuries [Illjustrirovannyj slovar' zabytyh i trudnyh slov iz proizvedenij russkoj literatury XVIII–XIX vekov], Orenburgskoe knizhnoe izdatel'stvo, Orenburg.
17. *Gomaa W. H., Fahmy A. A.* (2013), A Survey of Text Similarity Approaches, International Journal of Computer Applications, Vol. 68, № 13, pp. 13–18.
18. *Ilinskaya N. G.* (1989), Dictionary of uncommon and archaic words [Slovar' maloupotrebitel'nyh i ustarevshih slov], Sovetskaja Rossija, Moscow.
19. *Ivanov V. V., Solnyshkina M. I., Solovyev V. D.* (2018), Efficiency of text readability features in Russian academic texts, Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, Vol. 17, pp. 277–287.
20. *Jamieson S.* (2004), Likert scales: how to (ab)use them, Medical Education, 38(12), pp. 1217–1218.
21. *Khokhlova M. V.* (2016), Large Corpora and Frequency Nouns, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", pp. 237–250.
22. *Kilgarriff A. et al.* (2014), The Sketch Engine: Ten Years On, Lexicography, Vol 1, Iss. 1, pp. 7–36.
23. *Kilgarriff A., Rose T.* (1998), Measures for corpus similarity and homogeneity, Proceedings of the Third Conference on Empirical Methods for Natural Language Processing, Granada, Spain, pp. 46–52.
24. *Laposhina A. N.* (2017), Analysis of the relevant features for automatic readability assessment for texts in Russian as a foreign language [Analiz relevantnyh priznakov dlja avtomaticheskogo opredelenija slozhnosti russkogo teksta kak inostrannogo], Proceedings of the International, Proceedings of the International Conference "Dialogue 2017" [Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Bekasovo, available at http://www.dialog-21.ru/media/3993/laposhina.pdf.
25. *Leroy G., Kauchak D.* (2014), The effect of word familiarity on actual and perceived text difficulty, Journal of the American Medical Informatics Association, 21(e1), pp. e169-e172.
26. *Lüdeling A., Kytö M.* (eds.) (2009), Corpus Linguistics: An International Handbook, Vol. 2, De Gruyter Mouton, Berlin, Boston.
27. *Lyashevskaya O. N.* (2016), Corpus Instruments for Russian Grammar Studies [Korpusnye instrumenty v grammaticheskih issledovanijah russkogo jazyka], Jazyki slavjanskoj kul'tury, Moscow.

28. *Lyashevskaya O. N., Sharoff S. A.* (2009), The frequency dictionary of modern Russian language [Častotnyj slovar' sovremennogo russkogo jazyka], csv-version, available at: http://dict.ruslang.ru/freq.php.
29. *Piperski A. Ch.* (2018), Corpus Size and the Robustness of Measures of Corpus Distance, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018", pp. 578–589.
30. *Piperski, A.* (2017), Sum of Minimum Frequencies as a Measure of Corpus Similarity, Presented at the Corpus Linguistics 2017, Birmingham, available at https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper143.pdf.
31. *Reynolds R. J.* (2016), Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 289–300.
32. *Rogozhnikova R. P.* (ed.) (1997), Rare words in the works of the authors of XIX century: Dictionary-companion [Redkie slova v proizvedenijah avtorov XIX veka: Slovar'-spravochnik], Russkie slovari, Moscow.
33. *Russian National Corpus*, available at: http://www.ruscorpora.ru/new/.
34. *ruTenTen11*, available at: https://www.sketchengine.eu/rutenten-russian-corpus/.
35. *Schmitt N.* (2010), Researching vocabulary: a vocabulary research manual, Palgrave Macmillan, Basingstoke, UK.
36. *Schwarm S. E., Ostendorf M.* (2005), Reading level assessment using support vector machines and statistical language models, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05), pp. 523–530.
37. *Shaikevich A. Ya.* (2015), Measures of lexical similarity between frequency dictionaries [Mery leksicheskogo shodstva chastotnyh slovarej], Proceedings of the Internationational Conference "Corpus linguistics-2015" [Trudy mezhdunarodnoy nauchnoy konferentsii "Korpusnaya linguistica-2015"], Saint Petersburg, pp. 434–442.
38. *Sharoff S., Goldhahn D., Quasthoff U.* (2017), Frequency Dictionary: Russian, Quasthoff U., Fiedler S., Hallsteindóttir E. (eds.), Frequency Dictinaries 9, Leipziger Universitätsverlag.
39. *Sharoff S., Kurella S., Hartley A.* (2008), Seeking needles in the web haystack: Finding texts suitable for language learners, Proceedings of 8th Teaching and Language Corpora Conference (TaLC-8).
40. *Sharoff S., Umanskaya E., Wilson J.* (2013), A frequency dictionary of Russian: core vocabulary for learners, Routledge, NY.
41. *Sharoff S. A.* Frequency dictionary [Častotnyj slovar'], available at: http://www.artint.ru/projects/frqlist.php.
42. *Slioussar N. A.* (2005), Psycholinguistic survey in the structure of mental lexicon on the data of Russian verbs. [Psiholingvisticheskoe issledovanie struktury mental'nogo leksikona na materiale russkih glagolov], unpublished dissertation, Saint Petersburg.
43. *Solovyev V., Ivanov V., Solnyshkina M.* (2018), Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics, Journal of Intelligent & Fuzzy Systems, Vol. 34, № 5, pp. 3049–3058.

44. *Somov V. P.* (1996), Dictionary of rare and forgotten words [Slovar' redkih i zabytyh slov], VLADOS, Moscow.
45. *Taiga Corpus*. An open-source corpus for machine learning, available at https://tatianashavrina.github.io/taiga_site/.
46. *Van Heuven W. J. B., Mandera P., Keuleers, E., Brysbaert, M.* (2014), Subtlex-UK: A new and improved word frequency database for British English, Quarterly Journal of Experimental Psychology, 67, pp. 1176–1190.
47. *Winter B.* (2020), Statistics for Linguists: An Introduction Using R, Routledge, NY, London.
48. *Zhao Y., Jurafsky D.* (2009), The effect of lexical frequency and Lombard reflex on tone hyperarticulation, Journal of Phonetics, 37, pp. 231–247.