Decisions of Russian Constitutional Court: Lexical Complexity Analysis in Shallow Diachrony

Olga Blinova[1,2[0000-0002-5665-3495]], Sergei Belov[1[0000-0003-4935-9658]], Michael Revazov[1[0000-0003-2642-5587]]

[1] St. Petersburg State University, 7/9 Universitetskaya nab., 199034, St. Petersburg, Russia
{o.blinova, s.a.belov, m.revazov}@spbu.ru
[2] National Research University Higher School of Economics, 190068, St. Petersburg, Russia
ovblinova@hse.ru

Abstract. The paper is aimed at studying the texts of Russian Constitutional Court decisions, issued from 1992 to 2018. We analyzed the corpus, consisting of 584 decisions or 3,426,747 tokens (incl. punctuation marks) and tested the hypothesis about increasing lexical complexity of the documents. Using the R package *stylo* and MFW statistics, we got a picture that reflects the differences of the texts by years. The results of cluster analysis show that the texts of the 90s and 2000s are combined into the first large cluster. The second large cluster includes the texts of the 2010s. Using the R package *quanteda*, we obtained the values of 11 lexical diversity measures. We chose the index K (Yule's K) as a basic measure, relatively more reliable and independent of the text length, and then interpreted the values of this measure. In general, the value of K decreases over the years, except for the texts of 2006, in which there is a noticeable increase in the index value, and the texts of 1993, in which the outlier is observed. The calculation hapax proportion shows a picture of a gradual decrease in the share of hapaxes. If we apply the traditional approach to the interpretation of TTR values and derived metrics, we can conclude that, as the lexical diversity decreases and the proportion of hapaxes decreases, the texts become easier to read.

Keywords: Legal Linguistics, Decisions of the Russian Constitutional Court, Stylometric Analysis, Most Frequent Words, Lexical Complexity, Lexical Diversity, TTR, Yule's K, R packages, *stylo*, *quanteda*.

Introduction

This paper is aimed at studying the texts of Russian Constitutional Court decisions, issued from 1992 to 2018. The purpose is to verify the hypothesis, according to which the texts became more complex during the specified period (i.e. in shallow diachrony). At the moment, we were primarily interested in the lexical complexity.

The Constitutional Court is one of the youngest legal institutions in Russia. Its appearance in 1991 was associated with large-scale changes in the legal and political system caused by the rejection of Soviet legal and political system and the creation of a new democratic state in Russia.

The Constitutional Court, more than any other courts, was perceived as an “alien body” in the judicial system, since the Court was significantly different from all other judicial bodies in its objectives and duties. The task of the Constitutional Court was neither to solve a specific case, nor to draw conclusions about the rights and obligations of a particular citizen, but to compare the norms of a law challenged by a citizen and the provisions of the Constitution, ensuring the protection and implementation of constitutional principles.

The major function of the Constitutional Court is direct application and appropriate interpretation of the constitutional text. Citizens and legal entities’ complaints about the violation of their constitutional rights noticeably prevail among the cases considered by the Russian Constitutional Court. For example, in 2019, out of more than 3,500 judgments and rulings of the Constitutional Court, only three were made at the request of state bodies, about 30 were made at the request of the courts, and all the rest were made on complaints.

At the same time, it cannot be concluded that decisions of the Constitutional Court, initiated by citizens, are addressed specifically to these citizens. Of course, an applicant should understand whether the Constitutional Court supported her arguments as well as the outcome of the case consideration. However, only a small (operative) part of a Constitutional Court decision is devoted to this.

The rest is primarily addresses those bodies that must restore the violated rights, and not only and not so much the citizen who applied directly to the Constitutional Court, as those who find themselves in a similar situation. In this regard, the main addressees of Constitutional Court decisions are the legislative bodies, which should amend corresponding laws, and law-enforcement bodies (both executive and judiciary), which should interpret the legal provisions that they apply in accordance with decisions of the Constitutional Court.

However, it would be wrong to exclude citizens and legal entities from the addressees of the decisions. The Constitutional Court very rarely finds itself concluding the absolute, complete and unconditional contradiction of examined norms of the Constitution. More often, conclusions about unconstitutionality are made in relation to a particular interpretation of the impugned norm. Acts of the Constitutional Court become part of existing law, shall be applied along with statutes, the interpretation of which they strongly influence. Coming to court and demanding application of a provision, for example, establishing a social payment, citizens often must refer not only to the provisions of the law, but also to their constitutional interpretation in the practice of the Constitutional Court. That is why the **decisions of the Constitutional Court should be clear to all citizens**.

On the one hand, the field of activity for the Constitutional Court is an area of refined jurisprudence, free from description of factual circumstances, proving their existence and assessment of such evidence; on the other hand, it is a part of the existing legal regulation, along with statutes. The decisions of the Constitutional Court have the most significant difference with the latter, as these decisions are a result of the work of judges (i. e. professional lawyers). Many of the Constitutional Court judges are professors of law, others were appointed to the Constitutional Court after many years career in other courts.

A draft of each decision is prepared by a judge-rapporteur, while the final text of any decision becomes a result of the collective creativity of all judges and has no authorship.

That is why the assessment of the decisions' complexity is important and indicative. Such an assessment **demonstrates the ability of professional lawyers to be clear**, to write specialized texts addressing to a wide range of citizens, in an accessible manner.

1 This Paper's Structure and Recent Works

To test the hypothesis of texts becoming more complex with time, we used the capabilities of two R software packages (*stylo* and *quanteda*) [1], [2], [3]. Both packages allow to analyze non-structured text data.

Using the *stylo* package, we got a general picture that reflects the differences between the texts by year. Using the *quanteda* package, we received more detailed information about the lexical complexity of texts by different time periods.

The diachronic study of legal documents is the actively developing area. In particular, there are diachronic corpora of legal texts, for example, Corpus of Historical English Law Reports (CHELAR) [4]. A study of the texts of Russian legal documents in dynamics was carried out in [5], [6].

A research on the readability of texts of Constitutional Court is presented in [7]. The corpus of decision was analyzed using a simple readability metric, the Flesch-Kincaid formula, adapted for the Russian by I.V. Oborneva [8].

The Flesch-Kincaid formula for the Russian looks as follows:

$$\mathrm{FRE} = 206{,}836 - 60{,}1 \times \mathrm{ASL} - 1{,}3 \times \mathrm{ASW}, \quad (1)$$

where ASL is the average sentence length in words, and ASW is the average word length in syllables.

However, two points should be emphasized. Firstly, the coefficients of the Oborneva's formula were obtained by calculating the statistical characteristics of about 100 works of famous English-language literary classics (and translating these works into Russian). Thus, the formula **is not quite universal**, but is applicable primarily for the analysis of the complexity of texts of (translated) fiction; about the indicated problem, see [9]. Secondly, recent studies show that "**sentence and word length measures likely do not tap directly into linguistic components related to readability** … nor are they the only linguistic features related to readability" [10].

In this paper to assess the text complexity (lexical complexity) we also use the traditional method of assessing, calculating the TTR (type-token ratio), more precisely, we use a number of derived metrics.

TTR is the ratio of the number of unique tokens (types) to all document tokens. It is known, however, that the values of the TTR measure are not independent of text length, that is, documents of equal length should be compared to obtain relevant results. The solution to this problem can also be the use of derivative measures, see [11], [12]. Such measures are provided in the *quanteda* package.

In addition, to assess the lexical complexity, we use information on hapax richness and hapax proportion (the hapax is a token that appears in a sample once). Hapax richness is a measure that describes text from the same perspective as the TTR measure.

2 Methods for Assessing Text Complexity and Lexical Complexity

There is a rather long tradition of applying methods for assessing complexity (readability) to texts in Russian; for a review, see, for example, [13]. There is, among others, the traditional direction mentioned above, associated with the use of a wide variety of readability formulas. Only 5 or 7 readability formulas were adapted for Russian [14], [15]. So, the following metrics are used on the "LeStCor: Levelled Study Corpus of Russian" resource: Flesch-Kincaid grade level, Coleman Liau Index score, (Gunning) Fog, SMOG index, Automated Readability Index, New Dale Chall Adjusted Grade Level, Powers-Sumner-Kearl Grade Level [16].

To assess the lexical complexity, we can use information on:

- lexical density, the proportion of various content words in the texts;
- lexical richness and lexical diversity, measured by calculating the values of TTR or derived measures;
- number of words with abstract or concrete meaning;
- number of ambiguous words;
- number of function words (particularly prepositions);
- number of abbreviations

etc., see [17], [18], [19] and many others. TTR values well predict Russian text complexity, see [20].

3 Material

We analyzed a collection of judgments, consisting of 584 documents relating to the period since 1992, when the first judgment of the Court appeared. The distribution of decisions by year is described in Table 1.

The full texts of the decisions were taken from the database of the ConsultantPlus information system [21] and from the web-portal of the Constitutional Court [22].

There are no 1994 decisions in the text collection. This is due to the fact that the Constitutional Court suspended work at the end of 1993. The reason was the need to adopt a new law, regulating the Constitutional Court activities. As a result, in 1994 the Federal Constitutional Law "On the Constitutional Court of the Russian Federation" was adopted. The appointment of judges also took some time. In an updated form, the Constitutional Court resumed its work in 1995.

Table 1. The distribution of texts by year

Year	N of texts	Year	N of texts
1992	9	**2006**	10
1993	18	**2007**	14
1995	17	**2008**	11
1996	21	**2009**	20
1997	21	**2010**	21
1998	28	**2011**	30
1999	19	**2012**	34
2000	15	**2013**	30
2001	17	**2014**	33
2002	17	**2015**	34
2003	20	**2016**	28
2004	19	**2017**	40
2005	14	**2018**	44
Total	**584**		

We used the corpus, which contains texts combined by years. Accordingly, the corpus files received names like “1992”, “2003”, “2018”.

The text collection consists of 3,426,747 tokens (including punctuation), see Table 2.

Table 2. The description of text collection

Year	Tokens (incl. punctuation)	Tokens (mean)	Year	Tokens (incl. punctuation)	Tokens (mean)
1992	32489	3609.89	**2006**	62093	6209.30
1993	48072	2670.67	**2007**	93645	6688.93
1995	67961	3997.71	**2008**	56621	5147.36
1996	76767	3655.57	**2009**	97603	4880.15
1997	89138	4244.67	**2010**	145130	6910.95
1998	106561	3805.75	**2011**	185631	6187.70
1999	82536	4344.00	**2012**	229870	6760.88
2000	76945	5129.67	**2013**	239763	7992.10
2001	85747	5043.94	**2014**	217964	6604.97
2002	95908	5641.65	**2015**	239644	7048.35
2003	113715	5685.75	**2016**	214721	7668.61
2004	108004	5684.42	**2017**	293955	7348.88
2005	112454	8032.43	**2018**	253810	5768.41

4 MFW Statistics

4.1 Analysis Procedure in *stylo*

The *stylo* package was created for quantitative studies of writing style and can be used in authorship verification (including forensic linguistics) and diachronic studies, see [2].

We performed unsupervised multivariate analysis. Using the basic functions does not imply preprocessing and markup of the text collection (segmentation into sentences, lemmatization, etc.). We downloaded text data directly from the corpus files. Text metadata were included in the file names. Using the basic `stylo()` function, it is possible to analyze a corpus with the assistance of the following methods.

- Cluster analysis (CA), the results of which are visualized as a dendrogram, or a graph, showing the clustering of texts.
- Multidimensional scaling (MDS), as a result of which texts are displayed as ordered on the basis of several variables, so that similar texts are placed next to each other, and heterogeneous texts are separated, see [23, 19].
- Principal Component Analysis (PCA), which operates on the covariance between features (PCV) [Ibid, 933].
- Principal Component Analysis (PCA), which operates on the correlation coefficient matrix between features (PCR) [Ibid, 933].
- Building a Bootstrap Consensus Tree (BCT), summarizing various cluster analysis results based on the most frequent features occurrences and culling parameter values [2].

So, by means of the package one can find out how much the analyzed texts or text collections differ. As features for the analysis, n-gram sequences of tokens and characters can be used.

4.2 Analysis Results

At the stage of preprocessing, we performed tokenization and removal of stop words. When tokenizing, we used the built-in features of the package. To remove stop words, we took a stop word list from [24].[1]

The corpus size after the removal of stop words was 2,103,608 tokens. We formed a list of 1000 frequent features, then found the features that are used in at least 90% of the texts. In this way, we got a list of 1684 MWF, and then in the analysis we used 100 or from 100 to 1600 of them.

[1] We used a list of stop words to remove units that are not able to characterize the lexical peculiarity of a text or text collection. The list [24] consists of 159 high-frequency words and includes primarily function words, as well as some most common nouns, adverbs and verbs (*человек* ‘person’, *говорил* ‘said’ etc.).

Then we performed cluster analysis, multidimensional scaling, principal component analysis. As a measure of distance, where relevant, we used the **Eder's Delta measure**, which is recommended for highly inflected languages [25].

The results of the PCA, MDS, and CA (see Fig. 1 below) show that the texts of the 90s, 2000s, and 2010s can be described as two separate groups. One can see the following pattern: the texts of the 90s and 2000s are combined into the first large cluster. The second large cluster combines primarily the texts of the 2010s.

Thus, MFW statistics shows, that in general the texts before 2010 and after 2010 are clearly opposed, but the texts of 2005 and 2007 are adjacent to the group of texts of the 2010s. In addition, the texts of 1992 and 1993 are opposed to the rest of the texts written before 2010.

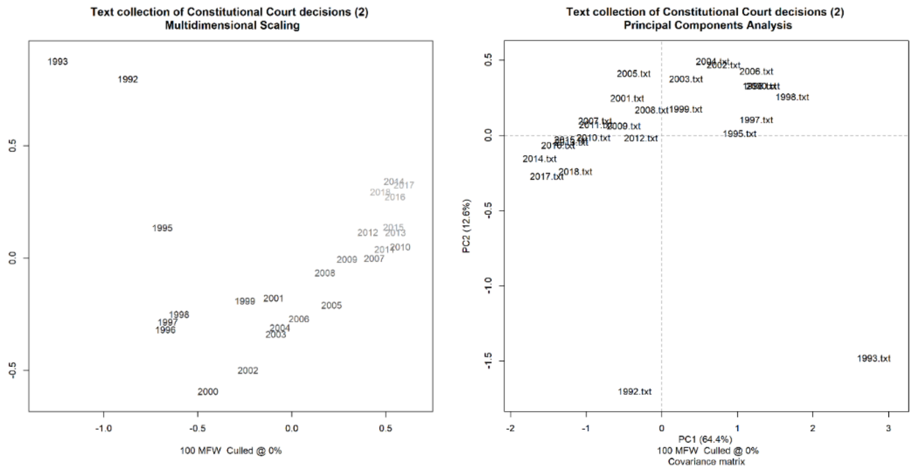

Fig. 1. MDS and PCA analysis results

On the whole, the results of calculating the Euclidean distance on normalized token frequency demonstrate a similar, but non identical patterns, see Fig. 2 (the dendrogram displaying normalized token frequency was obtained using *quanteda* package). The texts of 1992 and 1993 are contrasted with the rest of the texts in the corpus; in addition, the texts of 2012 fell into a large cluster containing the remaining texts of the 1990s and texts of the 2000s.

Finally, BCT allowed us to obtain the combined results of a cluster analysis (see Fig. 3).

5 Measures of Lexical Diversity

5.1 Analysis Procedure in *quanteda*

The *quanteda* package provides tools for a range of natural language processing tasks, see [3], it allows to perform tokenization, stemming, n-grams forming, selection and weighing of features [Ibid].

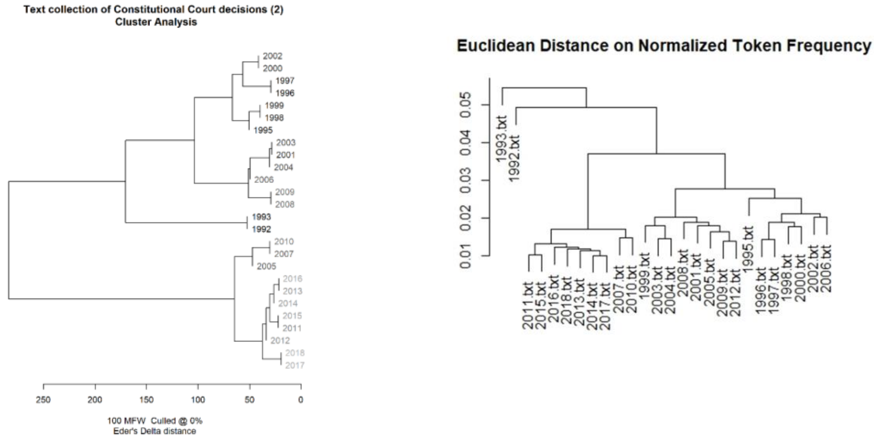

Fig. 2. Dendrogram of the results of CA using 100 MFW and normalized token frequency

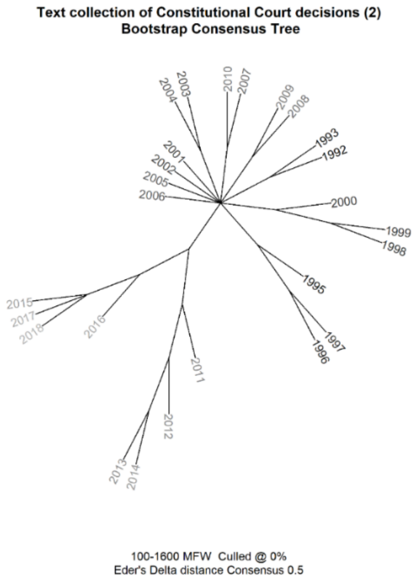

Fig. 3. BCT

We used the package capacities, related to the lexical diversity assessment. More specifically, we used TTR calculation and calculation of derived measures such as Herdan’s C (C), Guiraud’s Root TTR (R), Carroll’s Corrected TTR (CTTR), Dugast’s Uber Index (U), Summer’s index (S), Yule’s K (K), Herdan’s Vm (Vm), Maas’ indices (Maas, logV0, logeV0). The variables in all formulas are the number of types (V), the number of tokens (N), as well as fv (i,N), that is, the number of types occurring i times in a sample of length N [Ibid]. In addition, we calculated the amount and proportion of hapaxes.

When forming the corpus, we performed the removal of stop words, numbers and punctuation marks (since by default numbers were considered as tokens). The package uses the “Snowball” list of stop words [24].

5.2 Analysis Results

Using the package, we obtained the values of 11 measures of lexical diversity listed above.

Table 3. The values of lexical diversity measures

Year	TTR	C	R	CTTR	U	S	K	Vm	Maas	lgV_0
1992	0.21	0.84	30.67	21.69	27.80	0.88	47.29	0.07	0.19	6.81
1993	0.15	0.82	26.44	18.70	24.48	0.87	83.69	0.09	0.20	6.35
1995	0.17	0.83	34.44	24.35	27.64	0.88	49.08	0.07	0.19	6.95
1996	0.15	0.82	32.51	22.98	26.47	0.87	51.66	0.07	0.19	6.78
1997	0.14	0.82	31.77	22.46	25.88	0.87	50.32	0.07	0.20	6.71
1998	0.13	0.81	32.08	22.68	25.73	0.87	58.00	0.08	0.20	6.72
1999	0.15	0.82	33.35	23.58	26.73	0.88	41.15	0.06	0.19	6.84
2000	0.13	0.81	29.03	20.52	25.01	0.87	53.33	0.07	0.20	6.53
2001	0.16	0.83	35.65	25.20	27.64	0.88	32.22	0.06	0.19	7.00
2002	0.13	0.81	30.65	21.67	25.33	0.87	47.11	0.07	0.20	6.63
2003	0.14	0.82	35.70	25.24	27.02	0.87	40.67	0.06	0.19	6.96
2004	0.13	0.82	34.66	24.51	26.69	0.87	44.64	0.07	0.19	6.89
2005	0.14	0.82	37.21	26.31	27.58	0.88	33.63	0.06	0.19	7.05
2006	0.17	0.83	32.77	23.17	27.04	0.88	54.24	0.07	0.19	6.84
2007	0.15	0.83	36.50	25.81	27.70	0.88	29.59	0.05	0.19	7.03
2008	0.18	0.83	32.66	23.10	27.26	0.88	37.94	0.06	0.19	6.85
2009	0.14	0.82	33.82	23.91	26.56	0.87	32.82	0.06	0.19	6.85
2010	0.11	0.81	34.21	24.19	26.03	0.87	31.44	0.06	0.20	6.83
2011	0.11	0.81	36.55	25.85	26.49	0.87	29.70	0.05	0.19	6.96
2012	0.09	0.80	35.34	24.99	25.83	0.86	36.27	0.06	0.20	6.88
2013	0.10	0.80	37.45	26.48	26.44	0.87	26.67	0.05	0.19	7.00
2014	0.10	0.81	38.02	26.88	26.73	0.87	23.40	0.05	0.19	7.04
2015	0.10	0.81	38.25	27.05	26.66	0.87	26.71	0.05	0.19	7.04
2016	0.11	0.81	40.70	28.78	27.59	0.87	25.09	0.05	0.19	7.19
2017	0.09	0.80	39.84	28.17	26.88	0.87	24.46	0.05	0.19	7.12
2018	0.10	0.81	38.37	27.13	26.63	0.87	27.58	0.05	0.19	7.05

It is well known that the value of a simple TTR is affected by the text (or the sample) length, see for example [26], [27], [28], and many others. This problem can be solved in three ways.

1. It is possible to use samples of the same length.
2. It is possible to apply formulas with logarithms or with other transformations of variables N and V (Herdan’s C, Guiraud’s Root TTR, Carroll’s Corrected TTR, Dugast’s Uber Index, Maas’ indices).

3. It is possible to apply measures that make use of elements of the frequency spectrum (for example, the "Yule's K" measure), e.g. measures that take into account the number of hapax legomena (as in the "Honoré's R" measure) or hapax dislegomena, see [11], [12] for more details. We used the second and third possibilities, see Table 3.

The values of the indices TTR, C, S, K, Vm, Maas demonstrate a general decrease in time. The values of the indices R, CTTR, $\log V_0$ demonstrate a general increase in time. In general, the interpretation of the data is quite tricky, since the values of different measures are somewhat contradictory.

Therefore, based on the findings of [11], we chose the **K (Yule's K) index** as a basic measure, relatively more reliable and independent of the text length, and then interpreted the values of this particular measure.

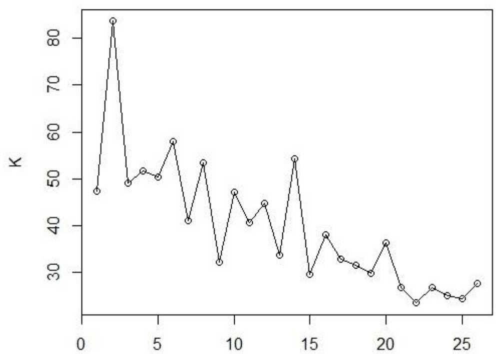

Fig. 4. The changes in the values of Yule's K (for 26 years)

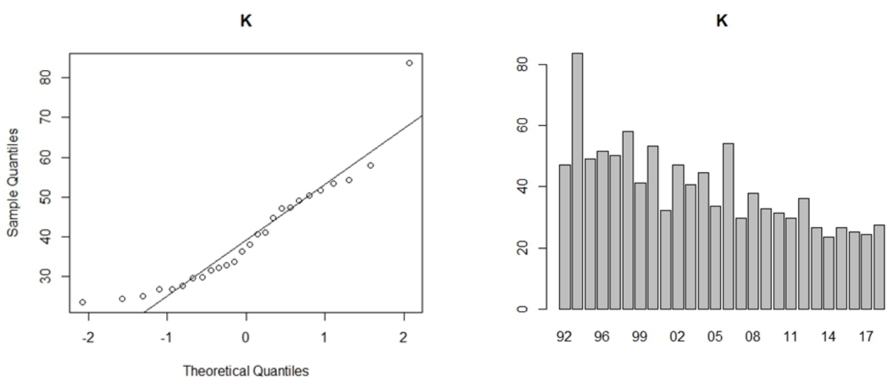

Fig. 5. Yule's K.

The value of the index K varies in the range from 23.40 to 83.69 (and the value of 83.69 observed in 1993 should be considered an outlier, see Fig. 4 and 5). In general, we can say that the value of K decreases over the years (except 2006 and 1993).

Accordingly, it can be argued that the lexical diversity of the texts of in time is decreasing.

The calculation of hapax richness (the number of tokens that appear in the sample only once) and the proportion of hapaxes show that the share of hapaxes gradually decreases, see Table 4, Fig. 6. However, the texts of 2006, 2008 and, to a lesser extent, 2007, 2005 and 2016 do not correspond to this general scheme.

Table 4. The number of hapaxes.

Year	N of Hapaxes	Year	N of Hapaxes
1992	2451	2006	3095
1993	2237	2007	4053
1995	3651	2008	2884
1996	3365	2009	3664
1997	3465	2010	4518
1998	3590	2011	5410
1999	3433	2012	5717
2000	2749	2013	6195
2001	3816	2014	5941
2002	3314	2015	6310
2003	4254	2016	6696
2004	4034	2017	7323
2005	4574	2018	6553

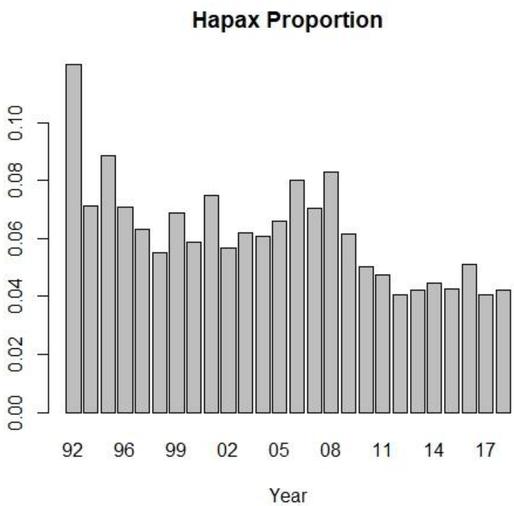

Fig. 6. The share of hapaxes

Thus, the share of hapaxes decreases over the years, the lexical diversity decreases over the years (texts have more and more repeating words). These two text evaluation options are easy to interpret in a consistent manner.

6 Conclusion

As a result of analyzing the corpus of Constitutional Court decisions we found out the following.

- MFW statistics shows that, in general, the texts before 2010 and after 2010 inclusive are clearly opposed.
- The texts of 1992 and 1993 are contrasted with all other texts (see, in particular, the clustering results after calculating the Euclidean distance on normalized token frequency). This can be explained by the fact that in 1994 the composition of the Constitutional Court was updated.
- The hapax proportion decreases over the years, the lexical diversity of texts also decreases.

If we apply the traditional approach to the interpretation of TTR values and derivative metrics, we can make a general conclusion that, since the lexical diversity is reduced and the proportion of hapaxes decreases, the texts become easier to read. Thus, our hypothesis of an increase in lexical complexity has not been confirmed.

There is an opposite approach to the interpretation of TTR values, we quote: "a lot of formal repetitions of the same words denoting legal entities and various legal terms interfere with the perception of the meaning of the sentence. In this case, we can say that reducing diversity not only does not simplify the text, but also causes the opposite effect" [5].

Legal texts use many repetitions. Though the presence of repetitions tires, it also allows to avoid problems with the interpretation of coreferential expressions. In addition, the process of text perception is affected by the priming effect (in particular, lexical priming).

Apparently, for the successful application of vocabulary-based measures for text complexity assessment, it is necessary to take into account at least some words' characteristics, that is, their semantics (first of all, abstractness/concreteness), their belonging to a certain part-of-speech class and general-language frequency.

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

References

1. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2018), https://www.R-project.org/.
2. Eder, M., Rybicki, J., Kestemont, M.: Stylometry with R: a package for computational text analysis. R Journal 8(1), 107-121 (2016).
3. Benoit, K, Watanabe, K, Wang, H., Nulty, P., Obeng, A., Muller, S., Matsuo, A.: quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30), 774 (2018). DOI: 10.21105/joss.00774.
4. Fanego, T., Rodríguez-Puente, P., José López-Couso, M., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C., Tamaredo, I.: The Corpus of Historical English Law Re-

ports 1535-1999 (CHELAR): A resource for analysing the development of English legal discourse. ICAME Journal, 41, 53-82 (2017).
5. Kuchakov, P., Saveliev, D.: The complexity of legal acts in Russia: Lexical and syntactic quality of texts: analytic note. European University at Saint Petersburg, St. Petersburg (2018). [in Russian].
6. Saveliev, D.A., Kuchakov, R.K.: Decisions of arbitration courts of Russian Federation: lexical and syntactic quality of texts, analytic note. European University at Saint Petersburg, St. Petersburg (2019). [in Russian].
7. Dmitrieva, A.V.: “The art of legal writing”: A quantitative analysis of Russian Constitutional Court rulings. Comparative Constitutional Review, 118 (3), 125-133 (2017). [in Russian].
8. Oborneva, I.V.: Automation of text perception quality assessment. Bulletin of the Moscow City Pedagogical University, 2, 221-233 (2005). [in Russian].
9. Solnyshkina, M., Ivanov, V., Solovyev, V.: Readability Formula for Russian Texts: A Modified Version. In: Batyrshin, I., Martínez-Villaseñor, M., Ponce Espinosa, H. (eds) Advances in Computational Intelligence. MICAI 2018. LNCS, vol. 11289. Springer, Cham (2018). DOI: 10.1007/978-3-030-04497-8_11.
10. Crossley, S.A., Skalicky, S., Dascalu, M.: Moving beyond classic readability formulas: new methods and new models. Journal of Research in Reading, 42(3-4), 541-561 (2019).
11. Tweedie, F.J., Baayen, R.H.: How Variable May a Constant Be? Measures of Lexical Richness in Perspective. Computers and the Humanities, 32(5), 323-352 (1998).
12. Oakes, M.: Corpus Linguistics and stylometry. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics: An International Handbook, vol. 2, Berlin, Mouton de Gruyter, pp. 1070-1091 (2009).
13. Reynolds, R.J.: Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, 289-300 (2016).
14. Begtin, I.V.: Readability.io [API for readability assessment of texts in Russian], https://github.com/ivbeg/readability.io/wiki/API.
15. Batinic, D., Birzer, S., Zinsmeister, H.: Creating an extensible, levelled study corpus of Russian. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), Bochum: Ruhr-Universität Bochum, pp. 38-43 [Bochumer Linguistische Arbeitsberichte 16] (2016).
16. LeStCor: Levelled Study Corpus of Russian, http://www.lestcor.com/calculator/.
17. Collins-Thompson, K.: Computational assessment of text readability: a survey of current and future research. In: François, Th., Bernhard D. (eds.) Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics, 165(2), pp. 97-135 (2014). DOI: 10.1075/itl.165.2.01col.
18. Evert, S., Wankerl, S., Nöth, E.: Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch. In: Proceedings of the Corpus Linguistics 2017 Conference, Birmingham, UK, (2017), http://www.stefan-evert.de/PUB/EvertWankerlNoeth2017.pdf.
19. Alfter, D., Volodina, E.: Towards Single Word Lexical Complexity Prediction. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 79-88 (2018). DOI: 10.18653/v1/W18-0508.
20. Ivanov, V.V., Solnyshkina, M.I., Solovyev V.D.: Efficiency of text readability features in Russian academic texts. In: Komp’juternaja Lingvistika i Intellektual’nye Tehnologii, 17, pp. 277-287 (2018).
21. Information-legal system “Consultant Plus”, http://www.consultant.ru/.

22. The Constitutional Court of the Russian Federation, http://www.ksrf.ru/.
23. Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., Nisbet, R.A. et al.: Practical text mining and statistical analysis for non-structured text data applications. Elsevier/Academic Press, Amsterdam (2012).
24. Snowball Stemming language and algorithms. http://snowball.tartarus.org/algorithms/russian/stop.txt.
25. Eder, M., Kestemont, M., & Rybicki, J.: Stylometry with R: a suite of tools. DH. (2013).
26. Mason, O.: Parameters of Collocation: The Word in the Centre of Gravity. In: Kirk, J.M. (ed.) English language research on computerised corpora; Corpora galore, 30, pp. 267-280 (2000).
27. Chipere, N., Malvern, D., Richards, B.: Using a corpus of children's writing to test a solution to the sample size problem affecting type-token ratios. In: Aston, G., Bernardini, S., Stewart, D. (eds.) Corpora and Language Learners [Studies in Corpus Linguistics 17], John Benjamins, Amsterdam, pp. 139-147 (2004). DOI: 10.1075/scl.17.10chi.
28. Kettunen, K.: Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? Journal of Quantitative Linguistics, 21, 3, 223-245 (2014). DOI: 10.1080/09296174.2014.911506.