Saint-Petersburg University

As a manuscript

Tarasov Nikita Andreevich

Hybrid neural network methods for analyzing the complexity of legal documents in Russian

Scientific specialty 2.3.1.
System analysis, management and information processing, statistics

Dissertation for the degree
of candidate of technical sciences

Translation from Russian

Scientific advisor:
Ph.D. tech. sciences
Blekanov Ivan Stanislavovich

Saint Petersburg — 2024

Table of contents

Introduction

The use of modern methods of collecting, processing and analyzing data leads to the development of existing areas and the creation of fundamentally new technologies in the information and legal sphere (LegalTech). Currently, LegalTech technologies, as a rule, include technological solutions that automate various legal processes: collection, processing and analysis of large volumes of legal data, information support for various legal processes, etc.

Automated processing of large arrays of legal texts using neural network models and technologies will make it possible to efficiently solve a number of problems of the legal process. In particular, modern language modeling methods can be used to solve problems of determining the complexity of legal documents that are relevant not only for individual companies, but also on a national scale. Thus, the efficiency of the legal process will increase by increasing the accessibility for large volumes of legal information.

Relevance of the research topic. Automatic processing of legal texts is of increasing scientific and practical interest. Modern data processing methods and artificial intelligence are significantly improving the quality of work with legal texts. The use of machine learning and natural language processing algorithms makes it possible to more effectively analyze, classify and interpret large volumes of legal information.

Intelligent methods for analyzing text data make it possible to both structure the content of individual documents and categorize corpora of documents of various types, taking into account semantics, as well as effectively identify features that describe various linguistic characteristics of the content. Methods based on machine learning algorithms and natural language processing technologies are capable of performing deeper analysis of texts and extracting semantically meaningful information from large document corpora.

The use of modern text processing methods in the legal field will help to significantly minimize the risks of errors when analyzing legal texts and ensure more accurate execution of legal procedures.

Large Language Models (LLM) are an effective modern approach for solving various problems in the field of intellectual text processing, including legal documents. However, for the effective use of such models for the Russian language and

taking into account various legal contexts, additional training and fine-tuning of the models is necessary. In programs presented in registration certificates [1—3] examples of the possibilities of additional training of language models and their adaptation for working with texts in Russian are shown. In order to create software systems methodological limitations of language models were examined in the context of the analysis of user content in social networks. The specificity of the language and non-standard document sizes unite the tasks of analyzing legal documents and user messages.

The complexity of legal documents often creates barriers to effective communication between different parties to the legal process. In this case, determining the understandability of documents is especially relevant for improving the quality of interaction between lawyers and non-professionals in the legal field. Thus, identifying unclear language constructions helps prevent possible double readings.

Thus, determining the complexity and understandability of legal documents makes it possible to increase the accessibility of legal information and allows to identify potential ambiguity and overly complicated parts of documents of various types - from agreements and statements to decrees and regulations.

The aim of this dissertation work is to develop and test methodological and instrumental tools for the intellectual processing of legal texts and provide algorithmic support for the process of determining their accessibility.

To achieve this it was necessary to solve the following **tasks**:

1. Study the current state of legal and linguistic research in the field of analysis of legal documents, identify current problems and determine possible methods for solving them.
2. Develop methodological approaches for collecting, processing and semantic analysis of the Russian legal language.
3. Develop a methodology for statistical assessment of frequency characteristics of legal language.
4. Identify and select linguistic characteristics of legal documents that best describe them in the context of complexity and readability.
5. To develop a software architecture for intellectual analysis of the complexity of legal documents based on hybrid neural network methods of using language models.
6. Conduct a comparative analysis of the complexity of documents of various substyles and genres using a hybrid complexity assessment model.

7. Conduct a practical analysis of the complexity of legal texts using the presented models and methods.

Scientific novelty of the research is as follows:

- Based on modern linguistic and legal research, as well as expert assessments, the most complete list of understandable characteristics of the Russian legal language has been identified and implemented.
- Based on modern scientific and technical methods, a system for intelligent data processing has been developed in the tasks of assessing the complexity and readability of legal texts.
- A set of approaches adapted for the Russian language has been developed, specialized databases of legal texts of various types and directions have been created.
- A hybrid neural network methodology for assessing the complexity of legal documents was presented and tested.
- A system for assessing the complexity of documents has been tested on various types of legal documents, both standardized (decrees, resolutions and other state legal documents) and in free form (answers to legal questions in the field of taxation).

Theoretical significance. The developed set of approaches and programs will significantly increase the efficiency of solving problems of intellectual analysis of legal documents related to the complexity and readability. The theoretical significance of the work is confirmed by participation in the following research projects:

- №19-18-00525 "Understanding official Russian: the legal and linguistic issues", 2020-2023 (Russian Science Foundation, participant)
- №96417361 "Legal and linguistic uncertainty in the texts of legal acts, their communicative features and legal functions," 2023-2024. (Government assignment - Grant for research at the expense of St. Petersburg State University, participant)
- №93825201 Project "Research Institute for Official Language Problems", 2022 (St. Petersburg State University, participant)
- №5-6-01/79 "Works on the study of the level of accessibility of perception of written responses from tax authorities to requests from individuals and organizations," 2023 (Federal Tax Service of Russia, participant)

- №92564627 “Center for International Media Research”, 2023 (Government assignment - Grant for research at the expense of St. Petersburg State University, participant)
- №16-18-10125-П “The distorting mirror of conflict: the role of network discussions in the representation and dynamics of ethnopolitical conflicts in Russia and abroad,” 2019-2020. (Russian Science Foundation, participant)
- №21-18-00454 “Mediatized communication and modern deliberative process”, 2023 (Russian Science Foundation, participant)

Practical significance. Based on the research conducted, a set of methods and programs has been developed for automated intelligent analysis of Russian-language legal texts in order to assess their complexity and readability. The proposed approaches and tools make it possible to analyze various types of legal documents, helping to accelerate the implementation of information technologies in real legal processes. The developed methods can find application in the scientific field (for example, in linguistics and law), as well as increase the efficiency of the work of specialists and improve the quality of interaction between the general population and government agencies.

Approbation of the work. The main results of the work were presented at the following conferences:

- 15th international conference SCSM 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Denmark, 23.07.2023
- International conference IAMCR Annual conference ’Inhabiting the planet: Challenges for media, communication and beyond’, France, 13.07.2023
- International fifteenth international scientific readings in Moscow “media and mass communications – 2023”: the era of uncertainty in modern media and journalism: challenges of big data and artificial intelligence, Russian Federation, 09.11.2023 - 10.11.2023
- 25-я International Conference on Human-Computer Interaction: HCI International - 2023 (’hybrid’ conference), Дания, 23.07.2023
- 27-й World Congress of Political Science (IPSA/AISP’2023), Argentina, 15.07.2023 - 19.07.2023
- International Conference “Dialogue 2022”, Russian Federation, 15.06.2022 - 18.06.2022

- Russian International Conference on Natural Sciences and Humanities with International Participation “Science of St. Petersburg State University – 2021”, Russian Federation, 28.12.2021 - 28.12.2021
- International Conference Networks in the Global World 2022, Russian Federation, 22.06.2022 - 24.06.2022
- International 13th Conference Social Computing and Social Media, SCSM 2021, held as part of the 23rd International Conference, HCI International 2021, Online, 24.07.2021 - 29.07.2021
- Corpus linguistics - 2021: international scientific conference, Russian Federation, 30.06.2021 - 03.07.2021
- 12th International Conference on Social Computing and Social Media, SCSM 2020, held as part of the 22nd International Conference on Human-Computer Interaction, HCII 2020, Denmark, 19.07.2020 - 24.07.2020
- 26th International Conference on Computational Linguistics and Intelligent Technologies “Dialogue”, Russian Federation, 17.06.2020 - 20.06.2020

Publications. The main results on the topic of the dissertation are presented in 9 printed publications, of which 4 — in periodical scientific journals indexed by Web of Science and Scopus [4—7], 5 — in abstracts of reports [8—12]. 3 certificates of state registration of computer programs were received [1—3].

Scope and structure of work. The dissertation consists of an introduction, 5 chapters and conclusion. The full volume of the dissertation is 114 pages, including 21 figures и 14 tables. The bibliography contains 156 items.

The introduction formulates the criteria, shows the relevance and novelty of the research, describes the theoretical and practical significance, and outlines the purpose and objectives of the research.

The first chapter describes the methodology for statistical assessment of the frequency characteristics of legal texts among various types of documents. Statistical data, the methodology for obtaining and processing them are important components of further analysis and create the basis for the descriptive characteristics of documents.

The second chapter provides a set of features that characterize legal documents according to the criteria of complexity, analyzes their effectiveness and proposes a method for solving the problem of complexity classification. Calculating language characteristics is the most common way to assess the understandability

of documents. Models based on these characteristics are further compared with algorithms built on the basis of language models.

Chapter three presents a hybrid complexity estimation method based on the joint application of language characteristics and large language models. The use of language models is a key element of the methodology. Their effectiveness in tasks of natural language analysis was evaluated, including in tasks of analyzing user texts - specific data with atypical vocabulary.

Chapter four provides a comparative analysis of the complexity of legal documents of various substyles and genres, based on the use of a hybrid semantic model for predicting complexity.

The fifth chapter provides an example of adapting the proposed methodology to solve the problem of analyzing answers to legal questions in the field of taxation.

In conclusion the results are summed up and the main results of the dissertation work are formulated.

Main scientific findings:

1. Formalization and development of a hybrid neural network model for assessing the complexity of legal texts. Presented in the work [4], see sections 2-6. (the method was developed personally by the author of the dissertation).
2. Methods for analyzing text data based on language models have been developed and adapted, see works [5; 6; 8—11] (the author of the dissertation developed methods and conducted computational experiments).
3. Text parameters have been identified that most effectively model the complexity of legal texts, see works [10; 12] (the author of the dissertation developed testing methods and conducted computational experiments).
4. The effectiveness of frequency zone modeling was assessed in the context of computing the complexity of texts, see work [11] (the author of the dissertation conducted computational experiments).
5. An analysis of the complexity of legal texts of various genres has been carried out, see work [7] (the author of the dissertation conducted computational experiments).
6. Software methods have been developed for adapting language models to solve problems of analyzing non-standard texts [1—3] (the author of the dissertation developed methods and software implementation).

Main provisions submitted for defense:

1. A complex of modern hybrid neural network methods based on large language models in informational support of legal processes.
2. Methodological approaches for collecting, statistical and semantic processing of legal texts from various sources.
3. Methodological foundations for adapting large language models in the task of determining the complexity of legal texts.
4. A set of programs for informational support of research and experimental work with Russian-language text data of legal processes, including components of collection, intellectual analysis and visualization.

Chapter 1. Modeling lemma frequency bands for lexical complexity assessment of Russian texts

1.1 Introduction

The section 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.

There is a fairly long tradition of applying readability assessment methods to texts in Russian; for a review see [13]. 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" [14], see also [15], 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 [16]. 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., [17]. It is clear that, with the exception of some special cases, the application of the familiarity criterion is difficult or impossible to operationalize without using word frequency information.

1.2 Word frequency as a parameter for text complexity assessing

According to [18], 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.

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 [19], it is precisely these features that showed the highest correlation with complexity. In [20] metrics based on lexical features (including word frequency, average frequency of nouns, etc.) are evaluated as reliable, see also [21; 22].

Frequency information can be applied in various ways. The average absolute word frequency or mean log frequency [23], the total frequency of content words [20] 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 [24].

Lemma frequency can be estimated using frequency dictionaries or representative corpora. This section is focused on the problem of the general-language frequency modeling based on data from large Russian corpora.

1.3 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, as measured by its relative frequency in a large representative corpus, or its presence/absence in a reference word list" [14]. 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., [25]. As stated in [26], a representative corpus "might contain roughly 90% conversation".

In [24] 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).

1.4 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, [27].

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 [24]. 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. The conventional threshold value for low-frequency words in a 100 million word corpus is 5 ipm (items per million) [28]. Different threshold values are also used for ranks. High-frequency units are the words with a rank up to 2,000 [29][60]; mid-frequency units are words with ranks from 2,000 to 8,000–9,000 [29][70]. Rare units in the New Frequency Dictionary of Russian are the lemmas with a rank of 10,000 and more [28][229]. The entire frequency list can be divided into quartiles (for example, in [30] words from the lower quartile of the ranked frequency list are considered as lowfrequency ones); percentiles can also be used for this purpose, see [31].

1.5 Frequency data sources

This chapter compares frequency lists derived from three large web corpora: ruTenTen11 [32; 33], Araneum Russicum III Maximum [34], [35], Taiga [36], and the New Frequency Dictionary of Russian (NFDR), based on data from Russian National Corpus [37], [38].

Frequency lists were obtained from the corpora sites or from corpora creators (see Table 1). The list of possible combinations is obtained using NFDR. For single-letter lemmas, a separate search was performed.

Table 1 — Frequency data sources

Corpus	Composition	Size	Number of lemmas	Analyzer
RNC (NFDR)	genre-balanced RNC subcorpus	91,982,416 graphic words	52,138 with more than 37 occurrences	Mystem
ruTenTen11	Internet: news and commercial sites, blogs, social media	18 billion tokens	457,473 lemmas with more than 5 occurrences	Treetagger
Araneum Russicum III Maximum	Internet: news and commercial sites, blogs, social media	15,961,200,372 words	8,893,947 lemmas with more than 5 occurrences	Treetagger
Taiga	Internet: the articles from literary magazines, naive poetry, news from popular news sites and other texts	near 5 billion words	2,988,610 lemmas with more than 1 occurrences	UDPipe

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 [39—41] and many others. Three measures were chosen that point 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).

Firstly, the rank correlation analysis was applied, 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.

Secondly, two measures of overlap, considered in [42] ("Coverage" and "Enrichment") were applied. The Coverage measure is calculated by the formula:

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

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:

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

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, following the [42]) the absolute frequency of 20 occurrences was used. This is the socalled "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, [43][818].

Thirdly, the measure "Sum of Minimum Frequencies" (SMF) was applied, proposed by A. Ya. Shaikevich in [44], see also [45]. SMF is calculated by the formula:

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

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.

1.7 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	ruTenTen	Taiga	NFDR	X/Y	ruTenTen	Taiga	NFDR
Araneum	0.033	0.081	0.223	Araneum	0.022	0.006	0.157
ruTenTen		0.071	0.828	ruTenTen		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 [42]. In other words, this is "a (very rough) measure of the extent to which X is 'substitutable' with Y". 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.

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.

Table 3 — Values of the measures of overlap, threshold = 20^{10}

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

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 considering 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	ruTenTen	Taiga	NFDR
Araneum	0.056	0.024	0.264
ruTenTen		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 observed in the pair NFDR—ruTenTen11 (SMF=0.756). As seen earlier that the rank correlation coefficients for this pair also take the largest 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, it was 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.

1.8 Comparison by frequency bands

For a more detailed comparison of frequency lists by different frequency bands, NFDR frequency list was divided and ranked into 4 equal parts, then, using the ranks, 4 random samples (containing 20 lemmas from each quartile) were formed . For each lemma of 4 random samples, the values of relative frequencies were assigned according to all the compared lists.

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 $\geqslant 1$ 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 highfrequency, mid-frequency, and low-frequency words. Meanwhile, it is useful to have a convenient way of assigning lemmas to certain frequency bands.

Therefore, (following Chen [24]) the approach from Van Heuven [46] was used, where a new "Zipf-value" measure of frequency is proposed. The value of this measure is calculated by the formula

$$Zipf - value = log_{10}(ipm \times 1000)$$

The measure has the following advantages:

– A logarithmic scale is used;
– The values are easy to interpret;
– The scale allows us to separate mid-frequency words from high-frequency and low-frequency ones;
– Zipf-values are easy to calculate if we know ipm values.

The discussed approach is not the only one possible. Sharoff [47] propose another logarithmic measure of the frequency "FClass", 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).

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

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

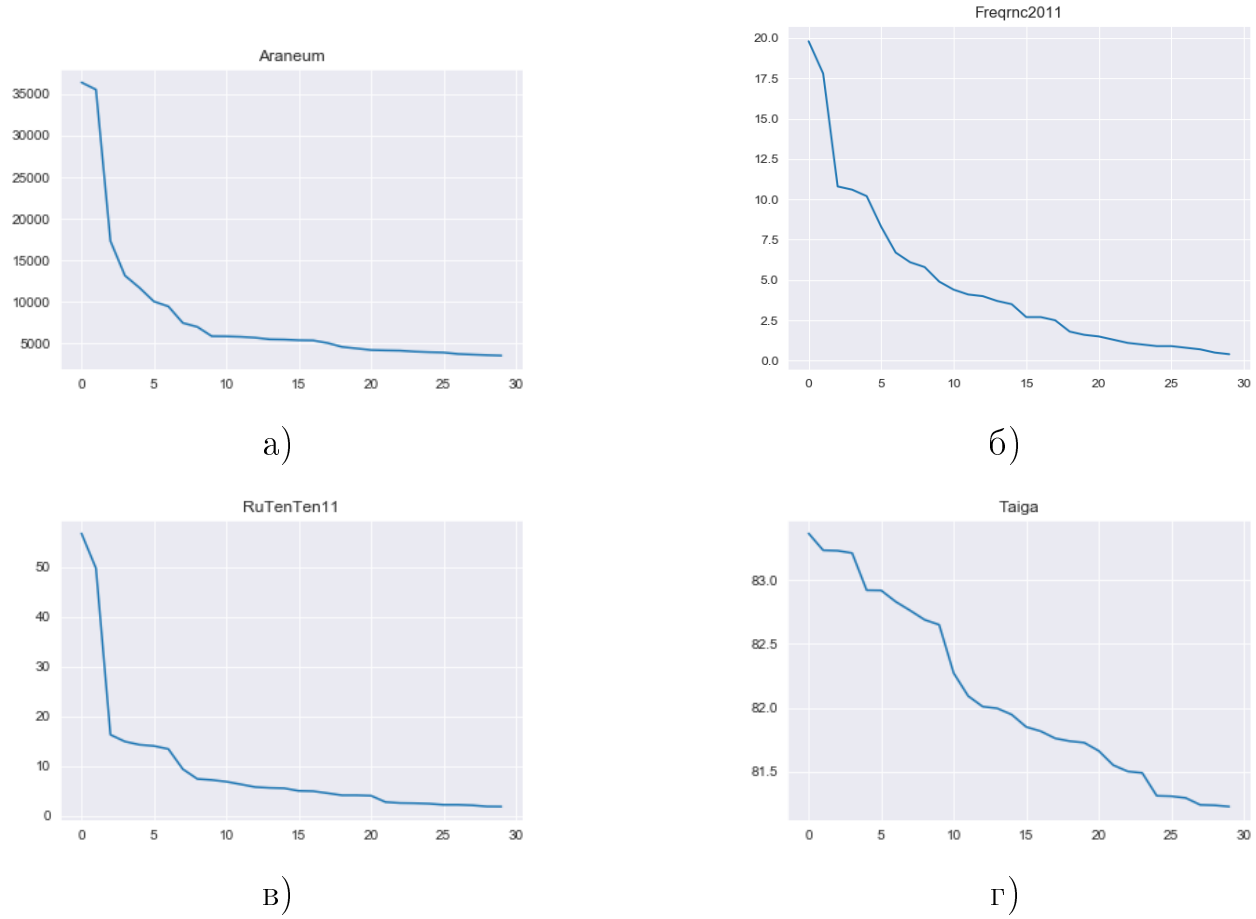
	$freq(w)$	MFW	$freq(max)$	FClass
NFDR	37	и "and"	3,293,765	16
Taiga	5	в "into"	11,076,749	21
Araneum	194	и "and"	563,822,183	21

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	22
Taiga	1	11,076,749	23
Aranuem	1	563,822,183	29
ruTenTen	1	503,894,565	29

The range of FClass values is greater than the range of Zipf-value. FClass scale does not look like a typical rating scale[48]. Accordingly, interpreting Zipf-values is a simpler task.



а) б)

в) г)

Figure 1.1 — Frequency distribution

Compared frequency lists, as shown below (see Fig. 1.1), obey exponential law. Therefore, Zipf-value can be used as a frequency measure.

It should be noted that lemmatizers assign different lemmas to the forms of Russian verbs, cf. превратиться (Pf)—превращаться (Impf), see Lyashevskaya [28] 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 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.

1.9 Chapter conclusions

In this section, the frequency lists derived from four Russian corpora were compared. The 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).

Chapter 2. Complexity metrics of Russian legal texts: selection, use, initial efficiency evaluation

2.1 Introduction

Legal texts, regardless of legal tradition and language, are characterized as complex, dark, confusing and incomprehensible to a non-lawyer, see [49; 50], [51; 52] and many more etc. This section is devoted to the description of a model developed to measure the objective complexity of legal texts in Russian. Model, based on 130 metrics, was developed taking into account the experience of research on linguistic complexity (including the complexity of legal texts), stylometric studies, as well as experimental work in the field of perception of legal texts.

The problems of determining the complexity of texts have been solved for quite a long time. In particular, there is a tradition of applying complexity assessment methods to Russian texts; for reviews, see, for example, [13], [53]. Along with the concept of "complexity", the concept of "readability" is used in the literature. Readability is understood as an assessment of the text obtained using parameters that are called latent in [54], in particular, readability formulas and measures of lexical diversity. Latent parameters are measurable, although not directly observable in the form of individual linguistic entities present in texts. Accordingly, complexity can be described as a more complex phenomenon; it is assessed both by referring to hidden parameters and using formal statistical (surface level) parameters [12].

2.2 Motivations for metrics selection

Complexity can be understood as a variable whose value is measurable for any (coherent) natural language text. Models for assessing complexity have evolved from simple (using readability formulas) to sophisticated (using a variety of metrics addressing vocabulary, morphology, syntax, unit frequency information, etc.).

The model also uses traditional complexity metrics; First of all, what has been said relates to the category of basic metrics and the category of "readability

formulas" (for more details, see Section 3). The experience gained in works on functional stylistics, stylometry and in psycholinguistic studies of perceptual complexity (difficulty) was also taken into account when developing this model.

Taking into account that the style of a text correlates with its complexity (that is, in some general case, business and scientific texts are more complex than news, journalistic, and conversational ones), the inclusion of style-specific metrics will be justified.

Let's give some examples. The studies mention the increase in the share of the noun, characteristic of business texts, and the fall in the share of the verb in the personal form. For example, in [55] it is stated that "the use of verb forms is reduced to a minimum in the official business style, which is distinguished by the most pronounced nominal character of speech," see also [56] and many others. etc. The increase in the proportion of nouns can be explained in different ways.

Firstly, it is customary to mention the frequency in official business style texts (hereinafter referred to as OBS) of "verb-nominal combinations" with "predicate splitting", see [55] and many others. etc., that is, about constructions with light verbs like 'to assist', 'to make replacement'. In the model, the shares of words of different parts of speech are counted, with the occurrences of constructions with light verbs are also taken into account.

Secondly, in the literature there is a judgment about the frequency of verbal nominalizations in OBS texts (regardless of their occurrence in constructions with light verbs). This feature in the model is taken into account in the word-formation metric and (partially) in the lexical metric, which takes into account occurrences of abstract words.

Thirdly, the increase in the proportion of nouns can be explained through the use in OBS of non-word term-like combinations such as 'tovarishchestvo sobstvennikov zhil'ya' (homeowners association) [57]. This feature is taken into account in the lexical metric, which counts occurrences of legal terms (including non single word ones).

Fourthly, the share of nouns is growing due to non-word derivative prepositions, the components of which are marked as nouns, e.g. 'v sootvetstvii' (in accordance), 'v svyazi' (in relation). This feature is also taken into account in lexical metrics. It appears that all four explanations are relevant. The example shows that taking stylistic work into account allows complexity to be analyzed in more detail. Practical stylistics recommends not to overuse passive constructions

in non-bookish styles, see [55], as well as [58] and many others. etc. Works on the perception of legal texts show that passive constructions are more difficult than active ones, see, for example, [59].

Accordingly, in the model, among the metrics of the category "individual grammes" there is a share of word forms in the instrumental case (since the instrumental case encodes the agent in passive constructions). In addition, among the syntactic metrics there is a proportion of occurrences of the passive subject of the main or dependent clause. Finally, the share of personal forms of the verb in -sya is taken into account, as well as (as part of part-verbal metrics) the share of full passive participles and the share of short passive participles.

It is important to note that experimental work on complexity (more precisely, perceptual difficulty) demonstrates that the diagnostic power of some traditional complexity metrics for measuring actual comprehensibility in experimental data is low. For example, [59] shows the low predictive power of readability formulas. It was also shown that the length of the sentence in the stimulus had little effect on how well subjects performed on the experimental paraphrasing task, and that sentences of the same length can vary greatly in actual comprehensibility.

Thus, a comparison of the findings of quantitative studies of text complexity and the findings of experimental studies allows us to look at the predictive power of complexity metrics more realistically. At the same time, the effectiveness of the metrics can be verified in testing.

2.3 Collection of metrics

The model uses 130 metrics to assess complexity, divided into the following categories:

1. basic metrics;
2. readability formulas;
3. shares of words of different part-speech classes;
4. frequency of lemmas;
5. word formation;
6. individual grammes;
7. lexical and semantic features, non-word expressions, hypertext links;

8. syntactic features;

9. coherency assessments.

The model uses 28 basic metrics. They can be divided into basic quantitative and basic lexical. The first ones are aimed primarily at measuring the length of words and sentences (cf. ASL - "average length of a sentence in words", ASW - "average length of a word form in syllables", S - "average number of sentences per 100 word forms", etc.). Basic lexical metrics involve the calculation of lexical diversity indices, as well as the calculation of hapax shares.

The model uses 5 readability formulas adapted for Russian: Flesch-Kincaid formula [60], SMOG, ARI, Dale-Chale index, Coleman-Liau index, see [61].

22 metrics that take into account the proportion of occurrences of words from different parts of speech, developed taking into account the differences between the tagging tools used in the model. For lemmatization, part-sentence and syntactic markup, UDPipe (model "ru-syntagrus") [62] was used. For the second layer of more detailed partial marking and morphological marking, pymorphy2 [63] was used. Under the influence of [64], the following were introduced into the model: an analyticity index (the ratio of the number of function words to the total number of words in the text); verb index; substantive index; adjectivity index; pronoun index; autosemantic index (the ratio of the number of meaningful words to the total number of words; all function words and pronouns are considered "insignificant"). In addition, the following are taken into account: the ratio of the number of nouns to the number of verbs; shares of coordinating and subordinating conjunctions; shares of full and short adjectives; shares of full and short participles; proportion of pronouns-nouns; shares of predicates, gerunds, infinitives; fractions of numerals; fraction of particles; the proportion of one-word prepositions, as well as the proportion of comparative forms.

13 metrics are introduced that address the representation of n-gram part-speech tags in texts. For the effectiveness of metrics that take into account part-speech compatibility, see, for example, [65]. It is worth commenting separately on bigrams of the form 'NOUN + NOUN', trigrams of the form 'NOUN + NOUN + NOUN' and bigrams of the form 'NOUN + NOUN,*gent'. Their use is aimed, among other things, at identifying noun phrases with several genitive arguments, which in the literature on stylistics are explicitly assessed as difficult to perceive, cf., for example, a quote from [55]: "The stringing of identical grammatical forms that make it difficult to perceive a text consistently depend on each other <...>. Epiphora often occurs when stringing together forms of the genitive case, which is

usually associated with the influence of an official business style" and the following example from the Budget Code of the Russian Federation: to ensure the necessary degree of confidentiality in the consideration of individual sections and subsections of federal budget expenditures and sources of financing the federal budget deficit, the State Duma approves the personal composition of workers groups <...>.

The "dynamic / static formula" proposed in [66] has been added, designed to separate texts that describe many events ("dynamic texts") from "static" texts. This metric contrasts business texts well with texts of other styles (official business style texts are more "static").

9 metrics were used, taking into account occurrences of lemmas with different general language frequencies, belonging to 9 frequency ranges. To calculate the values of this metric based on large Russian corpora, a consolidated frequency list of lemmas with Zipf value frequency indices was created, see [11]. The Zipf value in this list takes values from 0 (lowest frequency lemmas) to 8 (highest frequency lemmas). When assessing complexity, the proportion of occurrences of lemmas in the texts of each of the nine frequency ranges is taken into account.

To diagnose complexity under the influence of [57], one word-formation metric has been introduced. When calculating the values of this metric, the model refers to the level of lemmas, taking into account lemmas of the form *cia, *nie, *vie, *tie, *ist, *ism, *ura, *ische, *stvo, *ost, *ovka, *ator, *itor, *tel, *lnyi, *ovat (that is, counting the occurrences of some verbal and adjectival nouns, verbal adjectives and derived verbs). Note that the more complex cognitive processing of derived words compared to non-derived ones is confirmed in experiments on lexical decision making, see, for example, [67].

17 metrics of the category "individual grammemes" deserve a detailed discussion. The gender of nouns is taken into account, since abstract nouns used in legal texts are often neutral. The grammeme of the genitive case diagnoses complexity well, this is known from the literature on the issue, see, for example, [57]. The instrumental case encodes the agent in passive constructions. The set of personal forms of the verb is style-specific and genre-specific.

According to the literature on the issue, 3rd person forms are common in OBS, 2nd person forms are practically never found, and 1st person forms are used in a limited set of genres [55]. 11 metrics in the category "lexical and semantic features, non-word expressions, hypertext links" also address the described features of official business style texts. Among the category metrics: the share of text deixis means

that provide coherence; share of graphic abbreviations; share of abbreviations; part of the lemma 'to be'; share of legal terms; proportion of abstract lemmas; the share of lexical indicators of deontic possibility and necessity; proportion of non-word prepositions; the proportion of non-word phrases in the function of a conjunction or allied word; the share of constructions with light verbs, as well as the share of references to federal laws like '231-FZ' (the metric is designed to take into account hypertext connections).

The 21st syntactic metrics take into account:

- features describing the organization of individual syntactic groups (noun phrase – the proportion of adjectival modifiers of the name; verb group – the proportion of adverbial modifiers of the predicate); a feature describing occurrences of appositive noun phrases ("Appos");
- features showing the presence of composed series (whether they are composed clauses or homogeneous members of a sentence; feature "Cc", which describes allied means, as well as the sign "Conj", which describes the number of conjuncts, including those introduced without conjunction);
- features describing occurrences of sentential definitions (participles and participial phrases "Acl" and relative clauses "Acl:relcl"), sentential adverbials (gerunds and dependent clauses with personal forms of the verb, "Advcl"), various sentential additions ("Ccomp" , "Xcomp"), as well as so-called constructions with a sentential subject ("Csubj", "Csubj:pass"); units capable of introducing dependent clauses ("Mark") are taken into account separately;
- features describing occurrences of clauses with connective elements ("Cop");
- features that describe occurrences of passive constructions from different points of view ("Aux:pass", "Nsubj:pass", "Csubj:pass").

Finally, 2 coherence metrics assess the number of repetitions of nouns in neighboring sentences and the number of repetitions of tense and aspect grammes for verbs in the personal form (in neighboring sentences).

2.4 Model testing

To determine the quality of the selected 130 metrics and their ability to predict the complexity of texts, the following tests and comparisons were made:

– classification using the obtained metrics as parameters
– classification using language model vectors as parameters

2.4.1 Testing on the “plainrussian” dataset

The tests were carried out on the standard text dataset “plainrussian” by I. Begtin, which included texts divided into groups by level of education (from the 3rd grade of primary school to the 6th year of university)[61]. Due to the limited size of the test set (68 texts), the data for testing was divided into 3 classes: “simple texts” – up to 6th grade, “medium complexity texts” – from 6th to 11th grades, “complex texts” – texts level of higher education. The total number of documents for each group: “simple” – 14, “medium” – 32, “complex” – 22. XGBoost [68] was used as a test classification model.

2.4.2 Classification using language model vectors as parameters

A comparison was made with the USE (Universal Sentence Encoder) language model[69] using the modern “Transformer” neural network architecture, which has previously shown high efficiency in solving text classification problems [5; 9].. It allowed us to get an idea of the effectiveness of the selected metrics in the task of classification by complexity. In this way, the quality of coding the complexity of texts in the described approach was tested in comparison with the approach that encodes texts based on selected 130 metrics reflecting knowledge about natural language.

The model was tested with a preliminary split into test and training samples, followed by selection of hyperparameters using the “Hyperopt” library [70]. To select the parameters, 1000 models with different parameters were trained. The quality indicators cited above (see Table 7) are given for the optimized model using cross-validation[71] with data divided into 10 groups. This approach makes it possible to show the results more objectively and take into account the generalization of the model for previously unused data, which is especially important when working with small data sets.

Table 7 — Classification scores in the experiment with “plainrussian” dataset

USE кодировки			
Text type	Precision	Recall	F-measure
Simple text	0.506	0.583	0.524
Average text	0.667	0.333	0.419
Complex text	0.634	0.736	0.679
Кодировки метриками			
Text type	Precision	Recall	F-measure
Simple text	0.778	0.806	0.775
Average text	0.567	0.733	0.622
Complex text	0.849	0.778	0.811

Thus, the metrics allow obtaining more accurate estimates of the complexity of texts. “Complex texts” are identified most successfully, “simple texts” somewhat less successfully, and “texts of average complexity” least successfully.

2.4.3 Testing on a text set of social studies textbooks

The second iteration of the tests were carried out on a set of social studies textbooks, divided into groups according to classes of a general education school (grades 5 - 11)[72]. The data was also divided into 3 categories: “simpler texts” - grades 5, 6, 7, “texts of average complexity” - grades 8, 9, “more complex texts” - grades 10, 11. The total number of documents for each group: “simpler” – 5, “medium” – 4, “more complex” – 5, dataset size – 716 thousand words, average document length – approximately 1200 lines (one sentence per line).

All documents were randomly divided into fragments 100 lines long. The data was then labeled using UDPipe and pymorphy2, and 130 metrics were calculated for each fragment. After this, classification was performed. XGBoost [68] was used as a test classification model.

The final quality indicators for coding using metrics are shown in Table 8. In the experiments described, the effectiveness of 130 metrics was ascertained in the task of classification by complexity. Testing was carried out on data sets significantly different from target ones. Meanwhile, some metrics were specifically designed for

application to OBS texts. In texts of other styles, at least some of the features taken into account may describe rare or ultra-rare phenomena.

Table 8 — Classification scores in the textbook experiment

Encodings by metrics			
Text type	Precision	Recall	F-measure
Simple text	0.929	0.867	0.897
Average text	0.793	0.920	0.852
Complex text	0.971	0.895	0.932

2.4.4 Effectiveness of individual metrics

An experiment with "plainrussian" showed that 72 metrics are effective in the classification task. In an experiment with social studies textbooks, it turned out that for classification, what is primarily important is the Flesch-Kincaid formula, the coefficients (constants) of which were calculated precisely on the dataset with social studies textbooks by its creators[72], as well as 94 other characteristics.

The ten most effective metrics in the experiment with "plainrussian" included: the average length of a word form in letters, the proportion of full adjectives, the proportion of words with a length of 4 or more syllables, the proportion of word forms in the genitive case, the proportion of adjectives, the proportion of bigrams of noun and noun tags in genitive case, Flesch-Kincaid formula, the proportion of occurrences of the passive subject of the main or dependent clause, the dynamic / static formula and the average sentence length in syllables, see Fig. 2.1.

For the classification of textbook texts (see Fig. 2.2), the following metrics worked better than other metrics: readability formulas (FRE GL, SMOG, ARI), as well as the nominal vocabulary index, the proportion of inanimate nouns, the Colman-Liau index, the proportion of lemmas with "tails" like *cia, *nie, *vie, *tie, *ist (see Section 3 above about them), the proportion of full adjectives, the proportion of short adjectives and the proportion of adjectival modifiers of the name.

Figure 2.3 shows the metrics that were effective in both experiments. They are ranked by overall importance and selected as follows: the weight of each of the elements (i.e., metrics for a specific data set) does not exceed 70% of the total.

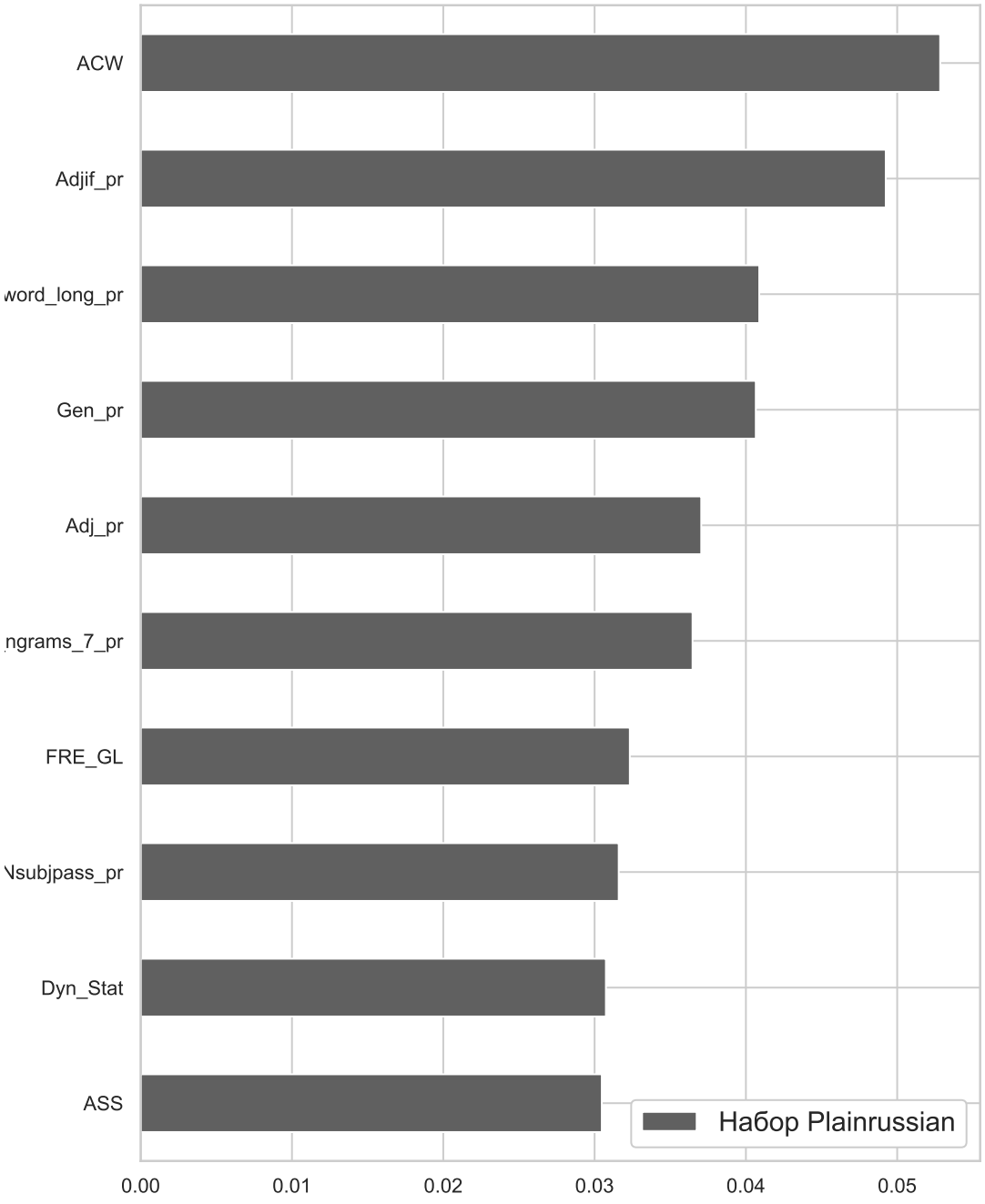


Figure 2.1 — Top-10 metrics, “plainrussian”.

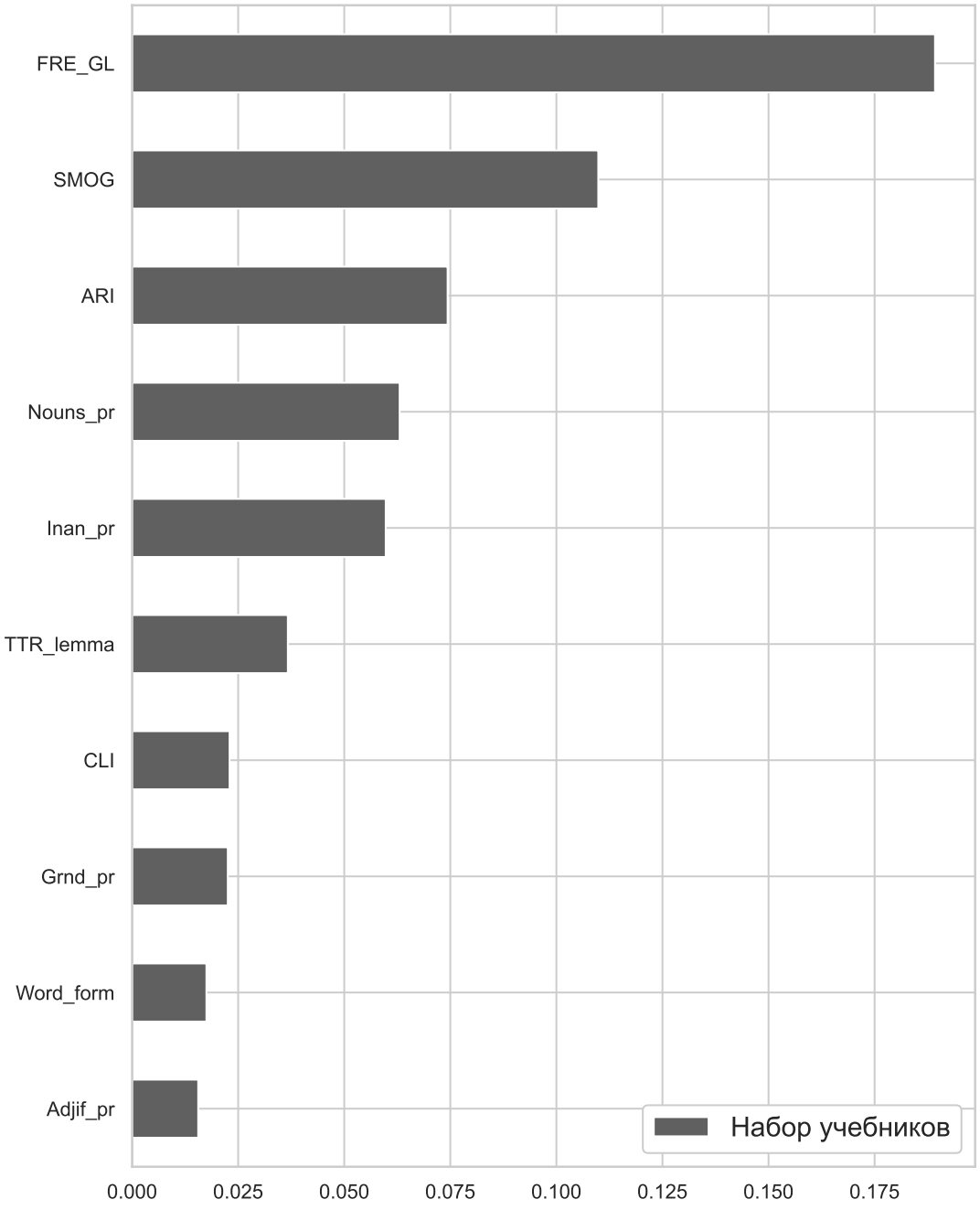


Figure 2.2 — Top-10 metrics, textbooks.

Among them (in descending order of importance): the Coleman-Liau index, the share of adjectival modifiers of the name, the share of lemmas with "tails", including certain word-forming affixes, the share of mid-frequency lemmas (Zipf value = 6), the verbosity index, the share of mid-frequency lemmas (Zipf value = 5), the proportion of noun pronouns, the proportion of constructions with light verbs, the number of word forms and the proportion of short participles.

2.5 Chapter conclusions

This chapter describes a model for assessing text complexity, which took into account 130 parameters, including style-specific ones (i.e., purposefully selected for Russian OBS texts). At the same time, the identified linguistic metrics show high efficiency in the task of representing texts with explicit linguistic parameters [8]. The continuation of the work will be focused on the transformation of the metrics-based model into a hybrid one. Using metrics in conjunction with effective language coding will allow complexity to be assessed both by linguistic parameters and by implicit features. When testing the model, a shortage of available Russian-language text sets with complexity (readability) ratings was apparent. "plainrussian" set was used, containing a total of 68 texts, as well as a significantly more extensive dataset of 14 textbooks [73]. Thus, testing was carried out on data sets that differ significantly from the target ones.

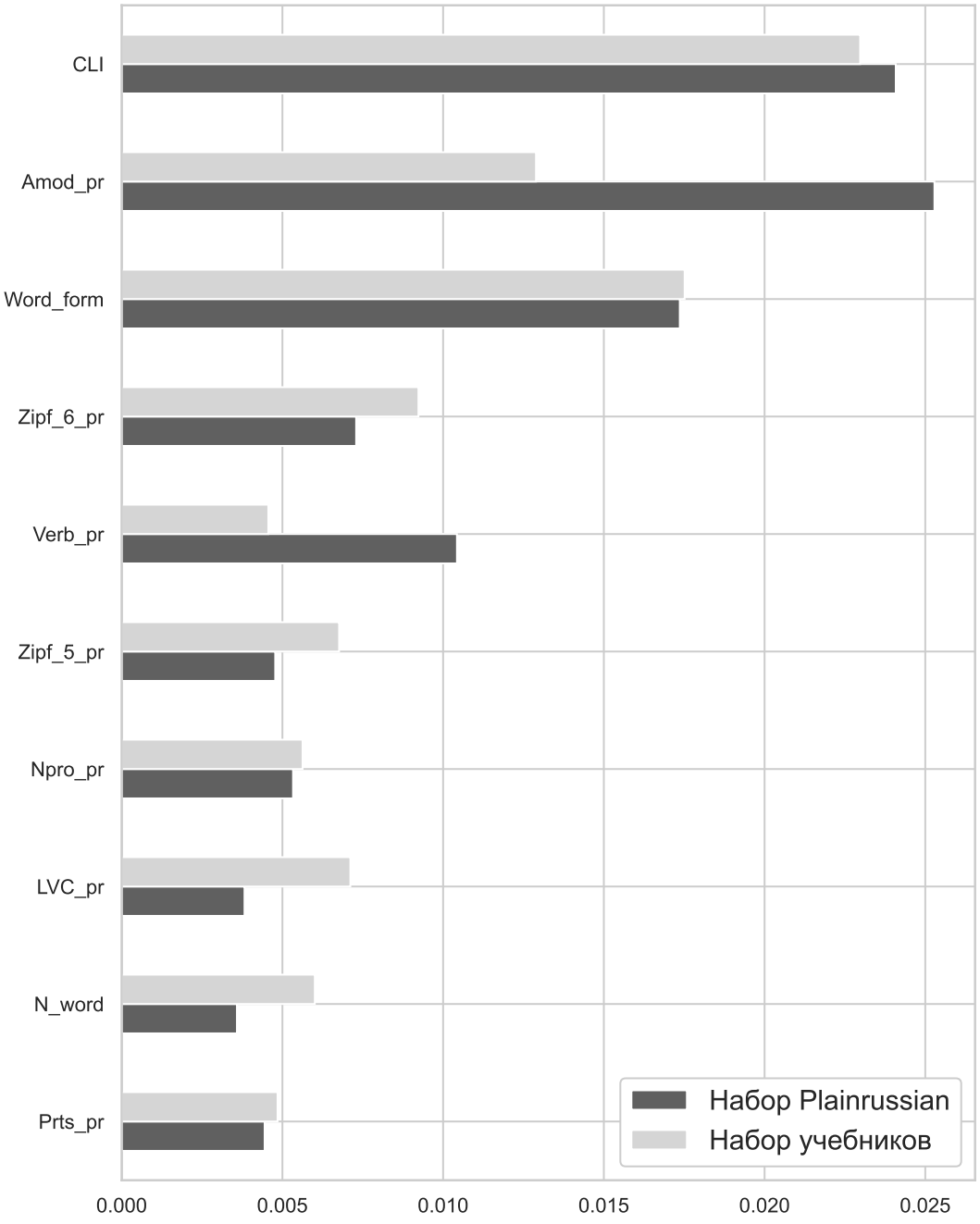


Figure 2.3 — Top 10 metrics by total importance

Chapter 3. A hybrid model of complexity estimation: evidence from Russian legal texts

3.1 Introduction

In this chapter the creation of a hybrid complexity estimation model is described which involves 130 metrics combined with neural network encodings. Linguistic features take into account lexical, semantic, syntactic properties of a text, its coherence, as well as sequences of part-of-speech tags, some word-formation patterns, and general-language frequency of lemmas. In addition, in-text references to other legal documents are considered (which is especially important when analyzing the laws).

The use of metrics in conjunction with efficient language coding allows one to estimate complexity from both linguistic parameters and implicit properties. The study [74] showed the success of such an approach in its most basic variation, i. e. adding neural network coding as a separate parameter for complexity estimation.

In terms of complexity, linguistic studies compare languages and dialects; language registers (or styles), and certain units (most notably words and sentences). The distinction between so-called "global" and "local" complexity is used [75]: the first branch of studies is interested in exploring languages "as such"; the second one measures complexity in particular linguistic subdomains and deals with phonological, morphological, syntactic, semantic, lexical and pragmatic complexity. The interlanguage comparison is dealt with by typologists (Ö. Dahl [76], J. Nichols [77] et al.), sociolinguists and contactologists (P. Trudgill [78], J. McWhorter [79] et al.). Perceptual complexity is studied by psycholinguists (see e.g. [80]). Computational linguists are also involved in complexity research, for an overview of approaches, see, for example, [81]. There is a rather long tradition of applying complexity assessment methods to Russian texts, for an overview see i. g. [13; 82].

The interest in the complexity of legal language is quite natural. Lingua legis has long been criticized for its verbosity, redundancy, lengthenings, syntactic overcomplication, archaic vocabulary, and unwarranted repetitions, see, e.g., [83; 84].

A number of studies are aimed at highlighting the characteristics of legal documents that cause their difficulty, at developing approaches to the "Plain language

movement", and the composition of recommendations for "Plain writing". Popular guides such as [85] give lawyers practical advice such as "omit surplus words", "use verbs to express action", "prefer the active voice", "use short sentences" etc. For the Russian research area, the problems associated with plain language have only been developed quite recently.

Russian legal texts have attracted the attention of complexity researchers, who, first, concentrated mainly on assessing legislative documents, and, second, used only readability formulas or other fairly simple and few measures.

For example, in [86] the texts of Constitutional Court decisions have been studied using a simple metric for assessing readability — the Flesch-Kincaid formula, adapted by I.V. Oborneva [87]. D. Saveliev and R. Kuchakov are also engaged in the study of complexity, see [88], [89]. In the cited papers, the authors have used only one lexical diversity measure (TTR, the value of which depends on the length of the text, hence the results of applying the metric may be questioned) and one syntactic measure ("Maximum Dependency Length", the distance between the head and the dependent on the dependency tree, calculated as follows "for each particular text one value is taken which is the maximum for all sentences of the text", (Ibid.)).

A new book [90] on the complexity of legislative texts identifies 9 factors, among them: "the share of verbs in the passive voice", "the share of verbs in relation to the total number of words in the text", "the average number of words in noun phrases", "the average number of participial clauses located in sentences after the word being defined, per sentence", "the average number of adverbial participle clauses per sentence", "the average number of words in sentences", "the average distance between dependent words in the sentence", "the average number of roots per sentence", "the average number of words per paragraph". Unfortunately, the authors (Ibid.) do not explicitly explain the reasons for their choice of parameters, which subsequently are not always clear to the reader. For example, it is not entirely clear what is meant by "the share of verbs in the passive voice", probably only the share of passive participles (since grammemes of the voice on the morphological markup layer are not assigned to the finite forms of the verb).

Thus, the authors of the studies on Russian legal language have focused on the complexity of legislative texts and the texts of judicial decisions. In addition, either only readability formulas or other, relatively few measures were used to estimate complexity.

In this chapter a complexity estimation model is proposed based on the combination of a variety of linguistic features and neural language model, trained on a large-scale data, and tested on three genre-diverse legal corpora. The goal is to test different machine learning models trained on a set of linguistic features and compare them to the results achieved by the deep learning approach. Here it can be hypothesized that a hybrid approach has the potential to achieve better quality than any individual model by utilizing both the explicit encodings of complexity measures and implicit representations of deep language model.

3.2 Related works

Recent developments in the field of natural language processing have presented new possibilities for feature engineering, and introduced new supervised and unsupervised methods for complexity estimation. In general, modern approaches can be split into two distinct categories: traditional machine learning approaches and deep learning models.

Classical machine learning approaches typically utilize a set of specific engineered features in conjunction with a classification algorithm. The introduction of classification models has made it possible to outperform traditional readability scores, such as the Flesch-Kincaid using unigram features and naive Bayes classifier [91]. Later feature sets have been expanded to include more sophisticated lexical, grammatical and discourse-based features [92]. [93] proposed a model for readability assessment for second language learners. The authors have utilized lexico-semantic features, parse tree features (such as grammatical relations), n-gram features and discourse-based features. The results have shown the effectiveness of these features and the SVM classifier. Similar results can be found in the research papers by [94] for texts in German language and [95] where the authors achieved the best results for the Italian language using a set of linguistic features in conjunction with the Random Forest classifier. [96] showed the effectiveness of linguistic features for the task of complexity assessment of the texts written by Russian learners of English. Authors compared a random forest classifier, k-neighbors classifier and logistic regression and have concluded that a random forest classifier with TF-IDF vectors

added as a feature obtains the best result. This result, in particular, shows the potential in combining the linguistic features and text encoding models.

Neural Network based approaches can be split into three general categories: general deep learning approaches (such as Feedforward Neural Networks - FNNs and Convolutional Neural Networks - CNNs), recurrent based networks - RNNs (including Long Short Term Memory - LSTM approaches [97]) and Transformer-based language models. [98] compared traditional machine algorithms with general deep learning approaches such as FNN and CNN. Neural network based approaches outperformed traditional ones such as random forests in most tests. Authors carried out the experiments on three datasets in Russian, collected from textbooks. [99] proposes a method of linking neural predictions of text complexity to linguistic properties of data.

Additionally, some models utilize neural encodings as their document representations, instead of traditional linguistic features, n-gram encodings or TF-IDF encodings. Word2vec [100], GloVe [101], FastText [102] are known to provide general high quality encodings. [103] compare these encoding techniques in conjunction with RNN to evaluate complexity in the Italian language. These approaches, howether, can be limiting in terms of application to a specific task. Transformer-based neural networks circumvent this issue by providing the opportunity to fine-tune the model to improve its effectiveness on a specific task. [104] discuss the applicability of the Transformer-based BERT [105] model for the task of readability assessment in German. Authors compare random forest regression with linguistic features, RNN based model with baseline BERT encodings and fine-tuned BERT for regression. The results show the effectiveness of the fine-tuned BERT model.

Thus, previous studies demonstrate the potential of both linguistic features and BERT embeddings. Different research works show inconclusive results on the subject of model choice for complexity assessment tasks — random forests classification and regression, RNNs and FNNs, SVM models all show the potential to achieve high quality results.

3.3 Data

Due to the lack of available supervised data on the topic of readability and complexity estimation in the Russian language for legal documents specifically, different datasets have been collected for the purposes of training and testing the model. Research on the complexity of Russian, in particular, commonly utilize textbooks data, see e. g. [106]. Thus, textbooks data is used for training to extract general patterns of text complexity for the language model. Additionally this data has been used to train the final hybrid model and estimate its quality. For final testing a set of legal documents has been used. These texts are used to test the effectiveness of the final model for the data, specifically related to the main task of this research — estimating the complexity of legal documents.

3.3.1 Training Data

Textbooks data was collected for the purposes of fine-tuning the Bert model and training the final hybrid model. The data consists of blocks of texts, randomly sampled from 1448 textbooks in the Russian language. Textbooks were split by paragraphs to obtain a large volume of training data and provide a language model with shortened texts. Textblocks size-limitation is important due to the fact that Transformer-based language models have a maximum input sequence length typically ranging from 128 to 1024 tokens. The data was additionally preprocessed: tables of contents, additional ending information and any non-textual information (tables, images etc.) were removed. Special symbols (excluding punctuation), occurring either naturally throughout the text or due to the errors of text file encodings were also removed. Training data was collected with variety and topicality in mind. Collected textbooks range in complexity from preschool and elementary school to high school and university books. Table 9 shows statistical features for the training data. Figure 3.1 shows the number of texts for each educational level ranging from 0 for pre-school level texts, 1–11 for years of school education and 12 for university level texts. Figure 3.2 shows the subjects and their corresponding amounts of texts.

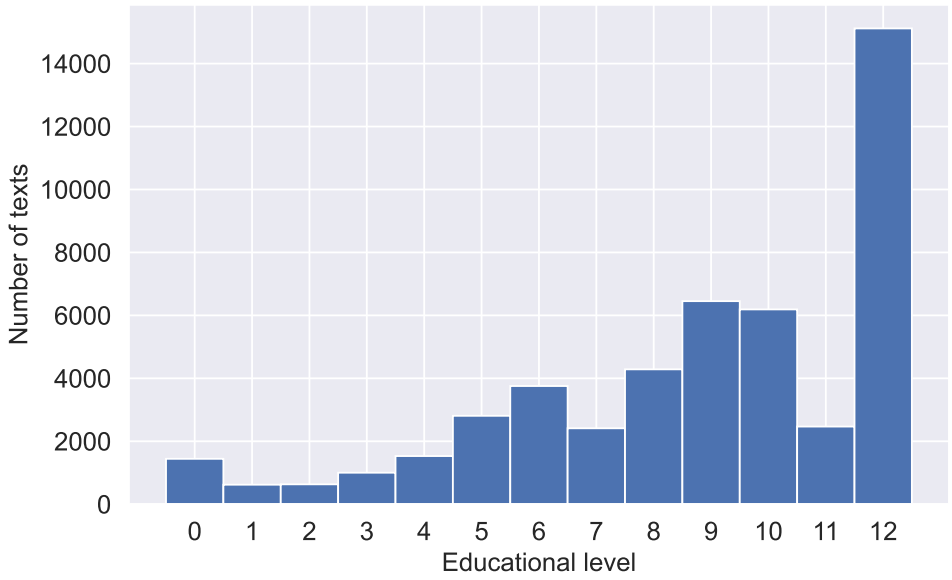


Figure 3.1 — Distribution of texts across educational levels with 0 for texts from pre-school books, 1–12 for schoolbooks and 12 for texts from university level books

The subjects were chosen due to expected similarities with legal documents (that is, the dataset includes textbooks on Jurisprudence, Social Sciences, Economics) and as capable of presenting samples of texts in Russian with varying levels of complexity (that is, the dataset includes textbooks on Literature, Culturology and History).

Table 9

	Total	Mean for each text block	Standard Deviation
Sentences	526 935	11	7
Tokens	9 939 730	204	151
Unique tokens	7 012 687	144	97

3.3.2 Testing Data

There is a significant number of Russian legal documents in the digital world; they are available, for example, through the legal information systems "ConsultantPlus"[107], "Garant" [108]. This makes it possible to create extensive corpora.

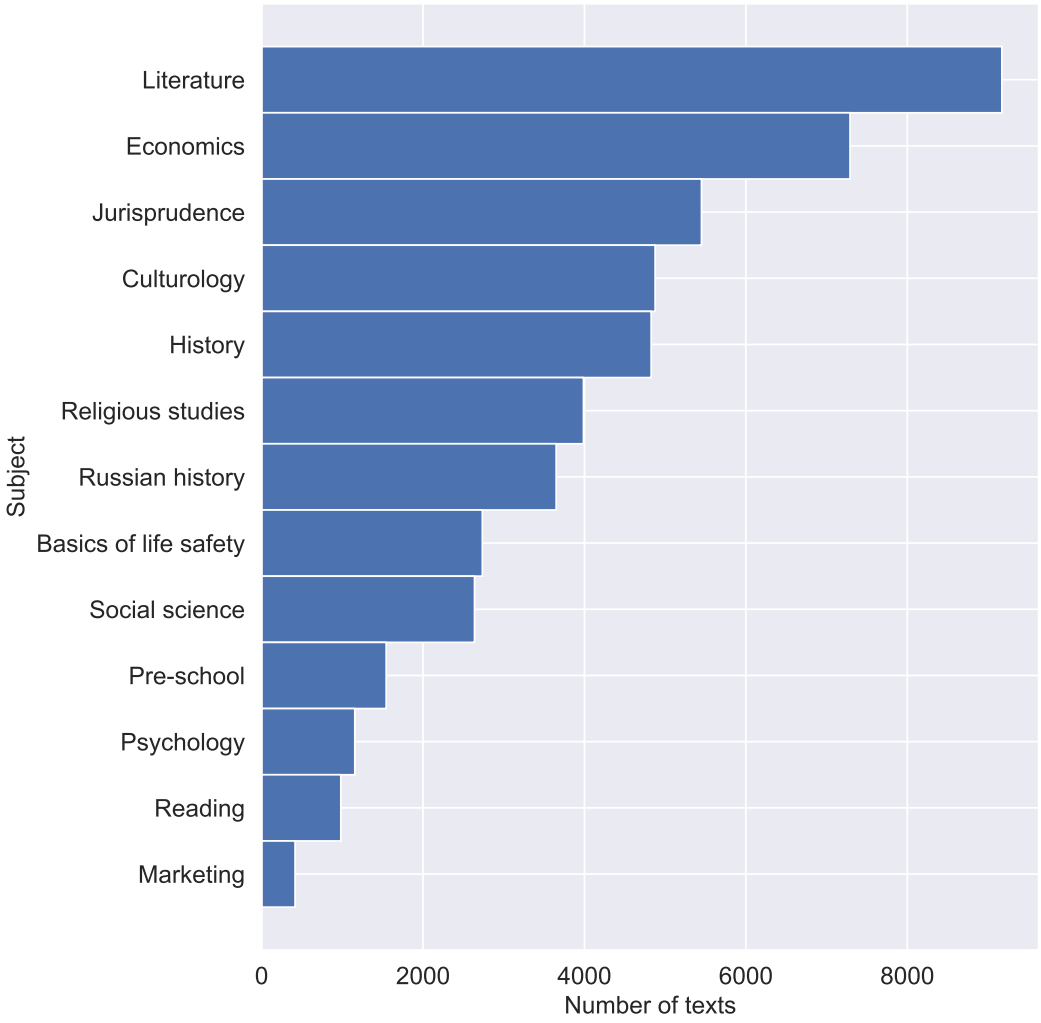


Figure 3.2 — Distribution of texts across subjects

The testing data are three legal corpora [10]. First, it is the "CorRIDA" corpus of Russian internal documents, consisting of 1,546 documents and containing 1,784 thousand tokens. Second, it is a corpus of decisions of the Constitutional Court of the Russian Federation "CorDec" of 3,427 thousand tokens, including 584 documents. Third, it is the "CorCodex" corpus of legislative documents, which contains 278 texts of codes, federal laws (a total of 3,227 thousand tokens).

Syntactic features are known to well predict textual complexity, see for example [109]. UD (Universal Dependencies) corpora have recently been increasingly used in assessing morphosyntactic complexity in both interlanguage comparison and comparison of text collections in the same language [110]. Therefore, UDPipe was chosen as the basic markup tool. As a tool for morphological analysis pymorphy2[111] was used. When choosing a pre-trained UDPipe model the accuracy statistics from [112] were used with the "russian-syntagrus" model being chosen.

After preprocessing, automatic lemmatization, morphological markup and syntactic parsing were performed. Each word form was assigned a double part-of-speech tag in terms of UDPipe and in terms of pymorphy2. The set of PoS tags of pymorphy2 allows, in particular, to distinguish between 'ADJF' (full forms of adjectives), 'ADJS' (short forms of adjectives), 'VERB' (finite forms of the verb), 'INFN' (infinitives), 'PRTF' (full form of participles), 'PRTS' (short form of participles) and 'GRND' (adverbial participles). This is convenient for assessing complexity, in particular, because there is a positive correlation between the number of full adjectives (as well as participles and adverbial participles) and complexity and a negative correlation between the number of finite verbs and complexity, see [113].

3.4 Linguistic features

To assess the complexity of Russian legal texts, 130 parameters were selected. The linguistic properties of Russian official texts (cf. the concept of "official-business style", oficial'no-delovoj stil'), described in research works on functional stylistics, as well as the features that are able to separate such texts from the texts of other styles when solving the problem of automatic classification by style, were taken into account.

All of the metrics used are conventionally divided into the following categories:

1. basic metrics;
2. readability formulas;
3. words of different part-of-speech classes;
4. n-grams of part-of-speech tags;
5. general-language frequency of lemmas;
6. word-formation patterns;
7. individual grammemes;
8. lexical and semantic features, multi-word expressions;
9. syntactic features;
10. cohesion assessments.

3.4.1 Basic metrics

The model provides the use of 28 basic metrics. Some of them are traditionally utilized in the tasks of classifying texts by complexity. All basic metrics can be divided into "basic quantitative" and "basic lexical" ones. The first ones are aimed, among other things, at taking into account the share of long words and long sentences ("long words" in the model are words consisting of 4 or more syllables). Basic lexical metrics imply calculating indexes of lexical diversity (simple TTR for word forms and lemmas; derived from TTR metrics "Yule's K" and "Yule's I", whose values do not depend on text length), and calculating the shares of hapaxes (hapax legomena and hapax dislegomena).

3.4.2 Readability formulas

The use of readability formulas is a common method of complexity estimation. It is now utilized in combination with other methods, see, for example, [114], and is embedded in a variety of textometric resources. The describing model uses five formulas: adapted Flesch-Kincaid formula [115], adapted SMOG (Simple Measure of Gobbledygook) formula, adapted formula for calculating the automated readability index ARI, Dale-Chale formula, Coleman-Liau index formula, see [116]. The formulas were adapted by Begtin using the text set which includes 68 documents categorized according to educational level (from the 3rd grade of elementary school to the 6th year of higher education).

3.4.3 Words of various part-of-speech classes

The metrics that take into account the shares of occurrences for words of various part-of-speech classes have been developed taking into account the differences between the markup tools used — UDPipe and pymorphy2, that is the differences

between the sets of PoS tags [117], [111]. Following [118], such indices were introduced into the model:

- "analyticity index" (the ratio of the number of function words to the total number of words);
- "verbality index" (the ratio of the number of verbs to the total number of words);
- "substantivity index" (the ratio of the number of nouns to the total number of words);
- "adjectivity index" (the ratio of the number of adjectives to the total number of words);
- "pronominality index" (the ratio of the number of pronouns to the total number of words);
- "autosemanticity index" (the ratio of the number of content words to the total number of words).

In addition, the ratio of the number of nouns to the number of verbs was used; the occurrences of short and full adjectives, short and full participles are considered separately.

3.4.4 Part-of-Speech N-grams

The information on n-grams of PoS tags was decided to involve for complexity analysis under the influence of studies on quantitative analysis of style [119], [120]. In (Ibid.) the so-called "dynamic/static formula" was proposed to separate "dynamic texts" describing a sequence of events from the "static" ones containing descriptive passages, for more details see e. g. [121]. This metric allows one to successfully distinguish official documents (they are more "static").

3.4.5 General-language frequency

In assessing complexity, it is customary to take into account the length of the words of the text and their "familiarity" to the reader. The "familiarity" can be

operationalized through the information on the general-language frequency of text lemmas. In the framework of the model for the accurate accounting of frequency data on the basis of large Russian corpora a frequency list was created. This list contains about 1 million lemmas distributed into 9 frequency bands using Zipf values, see about the method [11]. Complexity estimation model is able to calculate the proportion of lemmas belonging to each of the 9 frequency bands and to distinguish between high-frequency, medium-frequency, and low-frequency lemmas.

3.4.6 Word-formation patterns

Derived words formed with the help of affixes are generally longer than generating ones. In addition, derivatives are more complex morphologically. This complexity makes derived words more perceptually difficult, which is confirmed experimentally, see [122]. In the model, word-formation data is extracted from the level of lemmas, in each document the proportion of lemmas with endings of the type **cija*, **nie*, **vie*, **tie*, **ist*, **izm*, **ura*, **ishhe*, **stvo*, **ost'*, **ovka*, **ator*, **itor*, **tel'*, **l'nyj*, **ovat'* is calculated. This allows us to take into account the usage of deverbative and adjective-derived nouns, verb-derived adjectives and some derived verbs.

3.4.7 Grammemes

The model uses 17 metrics, taking into account, in particular: word forms in the genitive, instrumental, dative case, neuter nouns, 3rd person verbs, full and short forms of passive participles, and finite verb forms with *-sja*.

3.4.8 Lexical and semantic features, multi-word expressions

The list of features assessed through a layer of lemmas or word forms is as follows

- the proportion of text-deictic expressions like *nastojashhij* 'present', *nizhesledujushhij* 'following', *vysheupomjanutyj* 'aforementioned' etc.;
- the proportion of graphic abbreviations;
- the proportion of letter abbreviations;
- the proportion of legal terms;
- the proportion of abstract lemmas;
- the proportion of lexical indicators of deontic possibility and necessity like *zapreshhat'* 'to forbid', *protivopravnyj* 'wrongful', *nadlezhashhij* 'proper' etc.;
- the proportion of multi-word prepositions like *v sootvetstvii s* 'in conformance with';
- the proportion of multi-word expressions used as a conjunction or conjunctive word like *vvidu togo chto* 'due to the fact that', *vsledstvie chego* 'whereupon';
- the proportion of light verb constructions like *okazyvat' sodejstvie* 'to render assistance' *osushhestvljat' podgotovku* 'to conduct preparation';
- the proportion of in-text references to the legislative acts, in particular, federal laws like *231-FZ* 'Federal Law #31'.

To calculate the values of corresponding metrics, the set of user dictionaries is applied, that is, the value of the metric is calculated as the share of units that matched the unit from the dictionary.

3.4.9 Syntactic features

High syntactic complexity is a characteristic property of official texts. An extensive literature describes parameters for estimating sentence complexity, clausal complexity, and phrasal complexity. An up-to-date review is given in [123]. An influential research in this field is [124]. A large number of syntactic complexity measures have been used by [74].

In the Russian language, the signs of complexity are considered to be, first of all, participial and adverbial participle clauses, complex and compound sentences, see, for example, [125], [109]. It is clear that the possibilities of syntactic complexity analysis are limited by the parsing format. The model uses UDPipe for depen-

dency parsing (see section 3.1.2 above for details), utilizes 21 syntactic metrics and takes into account, among other features: noun clause modifiers, adverbial clause modifiers, various sentential complements.

3.4.10 Cohesion

To assess referential cohesion, the measure "Cohes_1" (the number of noun repetitions in neighboring sentences) has been used. In addition, the metric "Cohes_2" was utilized , which takes into account the number of repetitions of grammemes of tense and aspect for finite verbs (also in neighboring sentences).

At the end of the section, it is worth adding that some parameters of complexity estimation are not independent of each other, in particular, according to Zipf's law of abbreviation, word length correlates with word frequency, see for example [126]. In addition, the representation in texts of the various features listed above can have both positive and negative correlation with the target complexity.

3.5 Experimental setup

The resulting model consists of three main modules as shown in Figure 3.3. The training process is performed in two stages. In the first stage a Transformer-based BERT model is fine-tuned to obtain the initial complexity prediction for each text. The texts are additionally encoded using a set of metrics described in section 4. Initial complexity predictions from the language model and feature encodings from predefined metrics are combined and propagated to the final testing module — a choice between different regression and classification models.

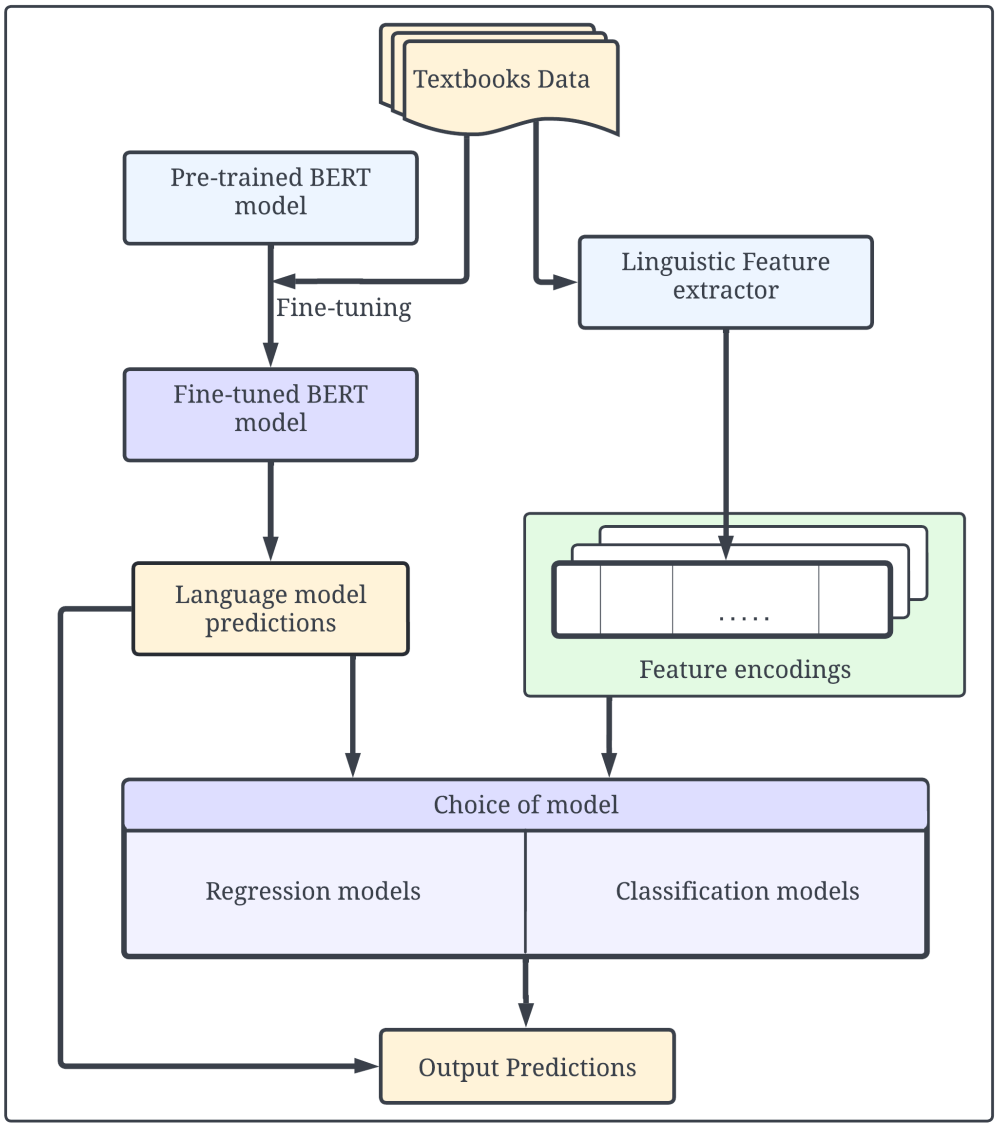


Figure 3.3 — Proposed training and testing pipeline including three main modules: Language model, Feature extractor and final hybrid model. The final model outputs both the result of neural model and the final result of the hybrid model.

3.5.1 Language model predictions

Transformer architecture has been utilized for a number of different natural language processing tasks both as a standalone approach and as part of a more complex combinational solutions. The basic idea of this approach lies in replacing recurrent layers with attention layers. This led to a significantly faster training process and better resource utilization due to parallelisation capabilities, previously impossible for recurrent networks and LSTMs. As such, Transformer is a fast and reliable method of language modeling that serves as a base for other more sophisticated and specialized algorithms. Bidirectional Encoder Representations from Transformer – BERT model improves on this idea by introducing the bidirectional architecture, introducing transfer learning procedure. Since its inception, transfer learning has become an integral part of most text analysis solutions. This approach consists of two main steps, i. e. initial pre-training of the model on a large scale and universal set of tasks (next sentence prediction and masked language modeling for BERT) and the fine-tuning step designed to adapt the model for a specific task.

The method of fine-tuning Transformer-based models pre-trained on a large scale data has been shown to provide high quality text representations across different NLP tasks. This process is done by adding an additional linear layer at the end of the pre-trained model and training it for a few epochs. The intuition behind this approach is that the initial pre-trained model learns generic language patterns, while the fine-tuning process allows the model to learn task-specific patterns [127].

In this research, a base version of RUBERT [128] was utilized, obtained from the Huggingface transformers library [129]. The model is pre-trained for the Russian language on the data obtained from various social media datasets. Initial pre-trained model consists of 12 layers, 768 hidden units per layer and 12 attention heads.

Due to the large number of categories for complexity in the dataset and their ordered nature it can be proposed that the regression approach could be more applicable. By defining the task as regression a potentially higher quality predictions in the corner cases can be achieved. Whereas classification predicts one of the outcomes without the context of their proximity to each other, the regression model can provide useful information by making predictions that lie closer to the real values even if not exact.

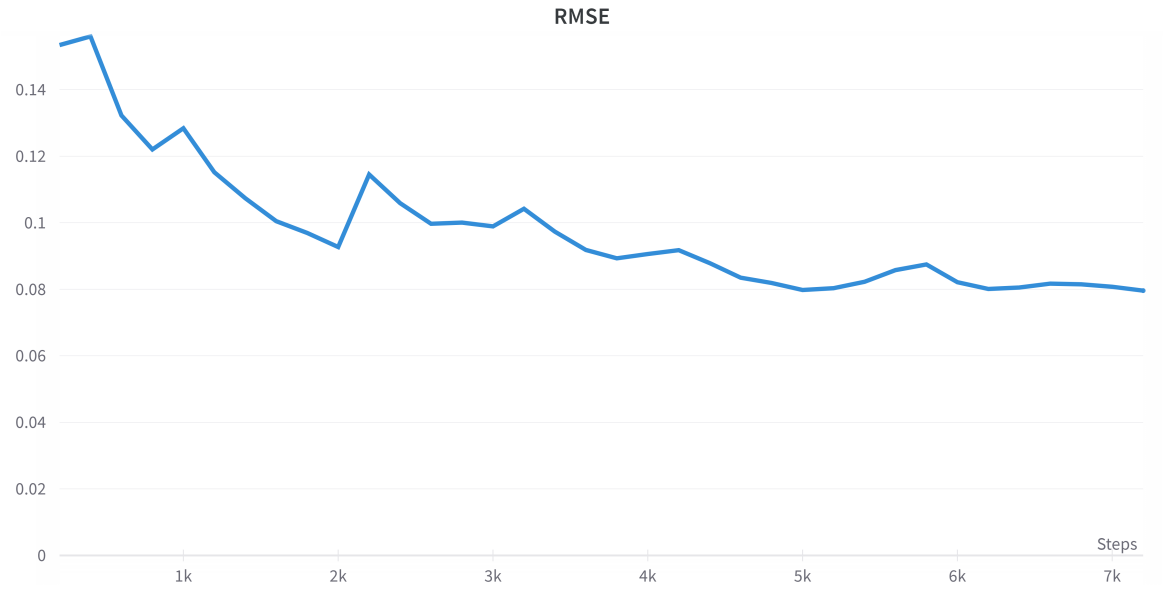


Figure 3.4 — Quality improvement during fine-tuning of the language model indicated by the RMSE metric

Our approach employs a standard fine-tuning process. It utilizes a pre-trained RUBERT tokenizer to split text blocks into tokens and add special padding and [CLS] tokens. Encodings are then passed through the model until the last layer where the hidden state of the [CLS] token is extracted and passed through a dense layer with hyperbolic tangent activation function. For fine-tuning we used AdamW optimizer [130] with 2e-5 learning rate, 16 batch size, 3 epochs, and 1e-2 weight decay. The model is optimized to find the best result in terms of RMSE loss for validation subset of data – 10% of the initial texts. Figure 3.4 shows the improvement of quality during the fine-tuning process.

3.5.2 Combining approach

To combine the linguistic features with the language model we obtain the output from the fine-tuned BERT model and use it as a feature in combination with linguistic features. This final vector representation is passed to another model. [74] utilize a SVM classifier for their choice of final model for its simplicity and frequent use in tasks involving adding numerical features.

Additionally the potential of other types of models is evaluated, including regression. With the large number of complexity classes (there are 13 categories in our case), there is a potential that regression models can provide a better result due to their ability to obtain a complexity score rather than direct class prediction. This can improve the quality and usability of the model. Whereas classification model can confuse between any class during the inference, regression model errors will still be close to the target value.

The quality of six models was tested: Linear regression, XGBoost [131] for regression, FNN for regression, SVM for classification, Random forest classification and XGBoost for classification. Linear regression and SVM classifier have been chosen to provide a baseline quality estimation using simple approaches. SVM classifier is also the model commonly utilized for complexity estimation task. The regression FNN model is a dense neural model which, in this case, consists of 3 hidden layers, 128 hidden units each. The model has been trained with Adam optimizer with 1e-3 learning rate. Random forest is a commonly used ensemble approach that trains a number of weaker decision trees on subsets of data and combines them into a stronger predictor, reducing the over-fitting. Extreme Gradient Boosting or XGBoost is a gradient-boosted decision tree (GBDT) machine learning library. It uses a technique where new models are introduced to correct the errors made by existing models. The hyperparameters for this algorithm were tuned using the Hyperopt library [70] to build 500 estimators for classification and regression tasks and find the set of optimal model parameters for each.

3.6 Experimental results

To compare the effectiveness of each method a set of metrics is used. Classification accuracy is measured as a basic percentage of correct predictions. For regression models this and all future classification metrics are defined by rounding the predictions to the closest category. Accuracy for university level texts (AUT) measures the accuracy of classification for texts with maximum complexity rating. It is measured to ensure the quality of predictions for texts of higher difficulty, presumably composing a large amount of legal text data. Precision, recall and f-measure are calculated using the weighted average of the values for each class. Root mean squared error is

Table 10 — Testing results showing the quality across different models and model combinations. Highlight indicates the best result for each metric.

	Accuracy	AUT	Precision	Recall	F1	RMSE	R2
Fine-tuned BERT	0.6308	0.9502	0.6366	0.6308	0.6311	0.0762	0.9173
Regression models							
Linear Regression with features	0.2095	0.2793	0.3821	0.2095	0.2333	0.1985	0.4399
Linear Regression combined	0.7053	**0.9873**	0.7163	0.7053	0.7028	0.0621	0.9451
XGBoost features	0.1491	0.2531	0.3871	0.1491	0.1378	0.2005	0.4283
XGBoost combined	0.5782	0.8055	0.6273	0.5782	0.5946	0.0728	0.9246
FNN with features	0.4918	0.8334	0.4834	0.4918	0.4839	0.1786	0.5465
FNN combined	0.7358	0.9741	0.7317	0.7358	0.7308	0.0654	0.9391
Classification models							
SVM features	0.3738	0.9455	0.3161	0.3738	0.2731	0.3226	-0.4787
SVM combined	0.3741	0.9462	0.3162	0.3741	0.2732	0.3226	-0.479
Random Forests with features	0.6002	0.9422	0.5952	0.6002	0.573	0.2179	0.3252
Random Forests combined	0.7775	0.9814	0.7814	0.7775	0.7723	0.0863	0.894
XGBoost features	0.6039	0.9137	0.5888	0.6039	0.5867	0.1968	0.4493
XGBoost combined	**0.7855**	0.9834	**0.7839**	**0.7855**	**0.7835**	**0.0605**	**0.9479**

measured to find the difference between predictions and true values in the regression problems. Lower values indicate higher quality. For classification algorithms the predictions are mapped to a 0 to 1 space. R2 score — coefficient of determination is a more straightforward regression score typically ranging from 0 to 1, however can be arbitrarily worse. Table 10 shows the results of testing for each model.

In all cases introduction of the BERT predictions provided an improvement in comparison with models trained only on linguistic features. In almost all cases the results were improved over the baseline BERT predictions. As highlighted in the table, XGBoost classification model trained on linguistic features and language model predictions achieved the best results on almost all metrics. This is true even

for regression-based metrics, indicating that incorrect predictions were close to the real scores. For regression models introduction of the language model predictions provided a more significant improvement in quality with the highest quality being achieved by the 3 layer neural network. Linear regression model with language model predictions achieved the best quality of predictions for university level text and obtained accurate predictions in general.

3.7 Discussion

The resulting model has been tested on the legal documents data. Initial predictions were obtained using the fine-tuned BERT model, combined with linguistic features and passed through the XGBoost model.

For "CorDec" dataset all documents were identified to have the highest complexity. For "CorCodex" data 95% of documents were given the maximum complexity score. "CorRIDA" data was found to be the most diverse with 83% of data identified as high complex documents. Figure 3.5 shows the distribution for the remaining files.

The observed differences between the three datasets are generally consistent with our expectations. The "CorRIDA" corpus of Russian internal documents and acts includes a little-studied category of legal texts, the so-called "internal documents". They are created in a particular state organization and regulate only the activity of this organization. The corpus contains documents addressed to the "ordinary citizen": to the applicant at the university, to a visitor at a museum or theater, to the patient at the clinic, etc. Apparently, it is primarily such official texts that we (i.e. Russian speakers who are not professional lawyers) periodically have some dealings with. For example, we sign "Consents to personal data processing", "Informed consents to medical intervention", or "Contracts for the provision of services". The internal documents are not always written by lawyers, standard templates are used to form them, but most importantly they are addressed to "ordinary speakers". Not surprisingly, the "CorRIDA" dataset consists not only of texts of maximum level of complexity.

The Constitutional Court Decisions, on the other hand, are written by highly professional lawyers, for a description see [132]. Such documents nominally are

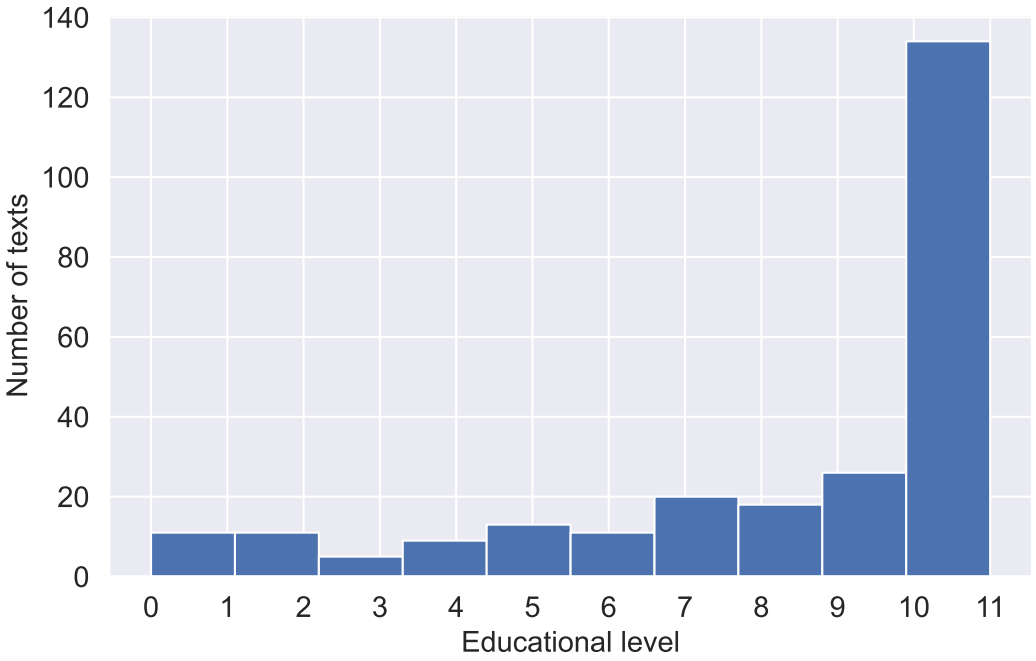


Figure 3.5 — Complexity distribution for CorRIDA data, excluding the university level texts

addressed to a wide range of citizens. However, lawyers themselves are concerned about the excessive complexity of the language of Constitutional Court decisions. Thus, [86] concludes that "the average judgment of the Court is written in too complicated language, aimed at a reader with a postgraduate education".

The third dataset (the "CorCodex" corpus) consists mainly of the texts of federal laws and codes. Complaints about the difficulty and incomprehensibility of the laws can be considered truisms, cf. the witty quote from [133]: "complaints about the excessive complexity of the law are as old as the law itself". Existing research works show that the complexity of legislative texts increases over the years, see [88]. Indeed, according to our results, only 11 of the 278 "CorCodex" corpus texts did not receive a score other than the maximum one, while 6 documents belong to the period from 1993 to 1999, 4 were written in the period from 2000 to 2003, 1 text was draft in 2010.

3.8 Chapter conclusions

In this chapter, a method of complexity prediction model hybridisation was proposed. A training dataset with texts was collected from textbooks in Russian with various levels of complexity on the subjects either related to the field of Jurisprudence or providing general language characteristics. The research demonstrates the effectiveness of BERT deep language model by itself and in combination with predefined linguistic features. The quality of models was measured on a set of metrics aimed to find the model, capable of high accuracy in general, high quality of predictions for complex texts in specific and low distance between predicted and actual values even in case of errors. These findings show that additional language model predictions provide a boost in quality for all regression and classification based models. The XGBoost model with tuned parameters, trained on features and language model predictions, has obtained the best result on training data and has been used in the final testing step. The additional tests on legal documents have showed the effectiveness of this approach in identifying complex texts, but have identified its biggest drawback, i. e. data dependence.

Chapter 4. Linguistic complexity of Russian legal substyles and genres

4.1 Introductcion

This chapter focuses on the linguistic complexity of legal sub-styles and genres in modern Russian. As pointed out by S. Goźdź-Roszkowski, "The expression "legal language" hides a multitude of specific classes of texts (genres) employed by various professional groups working in different legal contexts. Legal discourse spans a continuum from legislation enacted at different levels <...>, judicial decisions <...>, law reports, briefs, various contractual instruments, wills, power of attorney, etc. <...> through oral genres such as, for example, witness examination, jury summation, judge's summing-up, etc. <...>. This list is by no means exhaustive. It merely indicates the extraordinary diversity of legal discourse" [134]

Mattila et. al. [50] specifically points out that in some legal domains, some national legal traditions use "highly complex sentence constructions", scholarly vocabulary, formal and archaic language etc. Thus, legal genres can be characterized according to the level of linguistic complexity of the texts in question, see e.g. [135]) on two internationally used documents, [136] on contracts, [137] on different sub-varieties of legal language.

The purpose of this chapter is to find out the differences in linguistic complexity between legal documents, opposed by domain, sub-style and genre.

Approaches to classifying styles, sub-styles and genres, proposed by Russian functional stylistics are used. Legal texts are understood as a subset of the texts of "official business style" (rus. официально-деловой стиль).

Functional stylistics distinguishes legislative, justiciary and administrative sub-styles of the official business style. The first sub-style belongs to the sphere of legislation, the second one – to the sphere of justice, and the third one – to the sphere of administration, see, e.g., [56][329]. In addition, diplomatic sub-style is distinguished. The documents of this sub-style regulate legal relations between states.

Firstly, in this chapter, documents of national law and international legal documents are separated. This distinction is meaningful because many documents of international law are translated, i.e., linguistically, they may show significant differences from documents drafted in Russian.

Secondly, synchronous documents are considered. The notion of synchronicity is formalized as follows. "Synchronous" is generally considered to be all documents issued in the Russian Federation in 1991 and after (regardless of whether the documents are legally in force or not). Thus, documents of the Russian Federation are analyzed, but not of the USSR, not of the Russian Empire, not of Kievan Rus', etc. An exception to this definition of synchronicity are international documents in force, which (regardless of their date of issue) are also included in the analyzed Russian legal corpus.

Thirdly, particular legal genres are studied. Each of the sub-styles – legislative, justiciary, and administrative one – has a separate set of genres. At the same time, a variety of office and business documents related to accounting documentation, shipping documentation, etc., were not included in the set of documents of the administrative sub-style. Such documents are not included in the sample studied, because they obviously do not belong to the category of legal texts. For more information on the creation of the corpus of legal texts, a sample of which is analyzed in this chapter, see section 4.3.1 below.

Fourthly, only written legal genres are considered; oral genres remain outside the scope.

4.2 Literature review

4.2.1 Genre studies

In the Western linguistics, there are three main scholarly traditions for genre studies, namely rhetoric genre studies (RGS), systemic functional linguistics (SFL) and English for Specific Purposes (ESP), see e.g. Wang (2019). The first tradition understands genres as rhetorical actions, holding that "genre emerges from repeated social action in recurring situations which give rise to regularities in form and content" (Wang, 2019: 457). Genre studies within the new rhetoric approach focus more on the relationship between the text and the context than on the text features. SFL scholar J. Martin defines genre as "a staged, goal-oriented, purposeful activity in which speakers engage as members of our culture", respectively texts

with the same general purpose belong to the same genre [138][456]. Definition of genre in the ESP framework was proposed by J. Swales, who views the genre as "a class of communicative events, the members of which share some set of communicative purposes" [139][58].

Based on the ideas of the three genre theories, V. K. Bhatia proposed the following definition of genre: "Genre essentially refers to language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms by imposing constrains on the use of lexico-grammatical as well as discoursal resources" [140][27].

In addition to the genre itself as the main taxonomic unit, researchers use genre-unifying text category (super genre or macro-genre) and genre-splitting text category (sub-genre). Thus, when speaking of legal language, [50] proposes to distinguish legal sub-genres, according to the various sub-groups of legal authors (among which, in particular, judges, legislators, administrators, and advocates).

As pointed out in [141][13], "There is no fixed list of legal genres, even though a set of prominent legal text types can be identified. The core types include: 'legislative' documents (e.g. treaties, constitutions, statutes, statutory instruments, by-laws (sometimes 'bye-laws'), regulatory codes); 'private law' documents (e.g. contracts, orders, deeds, wills, leases, conveyances, mortgage documents, building contracts); and 'procedural' documents (e.g. opening speech in a trial, cross-examination, summing-up speech, jury direction)".

Active research into legal genres started in the 1980s, see [142][13]. There are research works on legislation and legal genres by Bhatia [140; 143], on lawyers' briefs by Kurzon [144], on contracts by Tiersma [145] and Trosborg [146], on legislative texts and contracts by Trosborg [147], on professional argumentation of lawyers by Howe [148], on apprenticeship into academic discourse community and degrees of linguistic intricacy by Iedema [149].

4.2.2 Complexity studies

There are plenty of research works related to the language complexity analysis, for an overview see e.g. [82]. The researchers of Russian-language legal documents

have focused on the complexity of texts of a particular type, or rather, even documents with a typical title issued by a particular institution, see Dmitrieva's work (2017)[86] on the complexity of Judgments of the RF Constitutional Court, and other research works, which will be discussed below. In the paper by Dmitrieva[86], complexity was evaluated using a single readability formula. Saveliev, Kuchakov [89] analyzed Judgments of the RF Subject's Arbitration Courts using two complexity metrics: simple TTR, whose value depends on text length, and Maximum Dependency Length, the distance from a head to its dependent on the syntactic dependency tree, calculated as follows: "for each specific text one value has been taken, which is the maximum for all sentences of text" [88]. At the same time, the authors interpreted TTR values in contradiction to the common approach, cf. the following quotation: "the multitude of formal repetitions of the same words, denoting subjects of law and various legal terms, interfere with the perception of the meaning of the sentence. In this case, we can say that the reduction of <lexical – O. B., N. T.> diversity not only does not lead to simplification of the text, but also causes the opposite effect" [88].

The most genre-diverse sample of Russian legal texts has been analyzed in [150]; in this research paper the author compares acts of the RF Constitutional Court, laws and codes, ministerial orders, and presidential edicts. Saveliev counts "the number of hard-to-read sentences" according to the "topic" of the texts (see e.g. the following topics: "Rules, instructions, directions, orders and other decisions", "joint-stock company", "Tsentral Bank of the Russian Federation", "Pension Fund of the Russian Federation"). In this case, the topics of the texts are not obtained as a result of their analysis, but according to the "General legal classifier of branches of legislation". Thus, the reader is not given a comparative analysis of genres or text types according to the complexity.

It can be summarized that the following categories of documents were considered for Russian in the context of complexity: legislative texts, i.e. laws [88; 90], and court judgments (see the research works cited above).

The most important thing is that the lawyers, engaged in studying texts of Russian legal domain, ignore genre distinctions as unconventional and irrelevant. That is, the authors were not interested at all in genre analysis and in the relationship between text genre and its complexity, as they applied other (legal, not genre-based) texts classifications, or did not apply any classifications at all. Meanwhile, it has been demonstrated that ignoring genre can significantly affect the

adequacy of the analysis of legal domain texts, see, for example, [151] on legal terminology. [152] showed that "readability assessment is strongly influenced by textual genre and for this reason a genre–oriented notion of readability is needed <...> with classification-based approaches to readability assessment reliable results can only be achieved with genre-specific models".

4.3 Materials and Methods

4.3.1 Legal documents

In order to understand which documents are to be included in the legal corpus, the taxonomies from the Russian legal databases and documentation databases were considered, namely Consultant Plus [107], Garant [108], Continent [153], Techexpert [154]. Based on this information, a preliminary list of document types was generated, containing 591 items (further – "list-591"). To evaluate this list, with assistance from legal experts an experiment was conducted in parallel annotation of document types by five assessors. The assessors (one Ph.D. and four Ph.D. students) went through the lines of the list and answered the question, "Is this <specific item on the list, type of document> a legal document or not?". The consistency of responses was assessed for each line (i.e., for each "type of document" separately), using a simple percentage of agreement. In this way a list of 108 "document types" correlated with written legal genres was obtained.

The next step in forming the list of genres was the analysis of dictionaries of legal terms Borisov [155] and Dodonov [156]. All the lines of the "list-591" (regardless of the lawyers' scores) containing "types of documents" were consecutively considered. Then the term corresponding to the document type was looked up in the dictionaries. Based on the interpretation of the term meaning, the decision was made to include the document type in the of genres to form the corpus. This procedure made it possible to identify the types of documents not mentioned in the "list-591" as well as to clarify the understanding of the genres in question. The following categories of documents were not to be included in the corpus of legal texts: "accounting documents" (e.g., advance report, audit report, balance sheet,

bill of lading), "payment documents" (e.g., debt claim, traveler's check, invoice), "foreign trade documents" (e.g., indent), "shipping documents" (e.g., bill of lading, goods release order), "cargo documents" (e.g., cargo receipt, cargo manifest, dock receipt, loading slip), "money documents" (e.g., cash voucher), "warehouse documents" (e.g., warehouse warrant).

The last stage of the list of document types formation was the analysis of (The Russian Classification of Management Documentation),[4] with the help of which the list of names of documents was expanded again. Combined list of legal "document types" (612 items) was then used to obtain the texts of documents from legal database sites and sites of state authorities.

4.3.2 Analyzing the data

Using the list of document types (see the previous section), legal documents were obtained and formed into a text collection. Then the names of documents were normalized from this text collection resulting in a list of genres, consisting of 306 items. All genres were divided into the following categories: international documents vs. documents of national law (administrative sub-style documents, legislative sub-style documents, and justiciary sub-style documents; further the corresponding documents will be referred to using acronyms ASSDs, LSSDs and JSSDs). In the next step selected genres were analyzed (a total of 68 genres, including 14 administrative, 24 legislative, and 30 justiciary ones). The basis for selection was the number of documents in a particular genre category and the public importance of the document (for example, the sample of LSSDs included the Constitution of the Russian Federation).

The lists of the analyzed genres of documents of national law are given in Table 11. The table also shows the number of genres considered (by sub-styles), the total number of documents of each sub-style, and the size of the samples in words.

SS	#Genres	List of Genres	#Docs	#Words
ASSDs	14	Ministerial Declaration of Goals and Objectives, Interaction Agreement, Ministerial Rules, Ministerial Agreement, Ministerial Minutes (Extract), Agreement on Information Interaction, Cooperation Agreement, Territorial Agreement, Performance Standard, Priority Project Change Request, Code of Ethics and Service Conduct, Ministerial Minutes, Ministerial Regulations, Ministerial Letter	938	3,798,795
LSSDs	24	RF Government Decree, Ministerial Order, RF Presidential Edict, Federal Law, Ministerial Decree, Labor Protection Instruction, Ministerial Instruction, RF Subject's Law, Ministerial Resolution, Ministerial Decision, RF Governmental Resolution, Regional Parliament Decree, Federal Parliament Decree, Sanitary Regulations and Standards, RF Law, RF Subject's Government Decree, Ruling Document, Ministerial Conclusive Statement, Labour Protection Rules, Ministerial Temporary Order, RF Instructional Letter, RF Code, RF Fundamentals of the Legislation, RF Constitution	14,813	58,430,223

JSSDs	30	Ruling of the RF Constitutional Court, Judgment of the RF Supreme Court, Ruling of the RF Supreme Court, Decree of the Arbitration Court of Appeal, Decree of the RF Supreme Court, Judgment of the City Arbitration Court, Decree of the RF Constitutional Court, Decree of the Federal Arbitration Court, Decree of the District Arbitration Court, Decree of the City Court, Decree of the Regional Court, Decree of the Appeal Court of general jurisdiction, Judgment of the Regional Arbitration Court, Decree of the Intellectual Property Court, Ruling of the Intellectual Property Court, Judgment of the Supreme Arbitration Court, Ruling of the RF Subject's Supreme Court, Verdict of the City Court, Verdict of the Regional Court, Decree of the RF Supreme Arbitration Court, Decree of the Regional Court, Decree of the RF Subject's Supreme Court, Prosecutor's of the RF Subject's Protest, Ruling of the Statutory Court, Conclusion of the RF Council of Judges, RF Supreme Court Protest, Ruling of the City Court, Decree of the Regional Arbitration Court, Ruling of the Regional Court, Verdict of the RF Subject's Supreme Court	26,436	50,138,771

Table 11 — Genres of National Legal Documents

The format of meta-labeling allows comparing documents of the same genre issued by different institutions, e.g. rulings of the RF Constitutional Court and rulings of the RF Supreme Court, RF Government Decrees and Ministerial Decrees. The International Law dataset consists of 1,617 texts, 6,400,239 words, includes international agreements, conventions, decrees, and judgments of international courts.

4.3.3 Complexity estimation model

Complexity model is described in detail in the previous chapter. The model has been composed in two main stages [4].

The first stage consists of complexity prediction, using a pre-trained Transformer based model. Transformer models have been proven to be effective at solving a wide array of language processing tasks using the idea of pre-training – initialization procedure aimed at capturing the core language features and fine-tuning – a process aimed at adapting the model for solving any given task. In this case RuBERT was chosen as a baseline pre-trained language model. An auxiliary dataset was collected for the purposes of fine-tuning the language model.

This dataset consists of text fragments, randomly sampled from 1,448 textbooks with complexity ranging from pre-school (used to describe 0 level of complexity), school textbooks of all grades (complexity from "1" to "11") and university level textbooks (describing the maximum level of complexity – "12"). The data contains fragments from the books on the subjects of Jurisprudence, Social Studies, Economics, Culturology, History etc. The subjects were chosen on the basis of being either good general language descriptors or their relation in this research area.

The decision to train the model using the textbook data was dictated by the lack of training data, designed specifically for legal texts. As such the textbooks on the topics, related to Jurisprudence, Economics and other social sciences have been chosen as the closest alternative. This solution can result in a more generalized complexity model. This model is capable of working across a wide range of data in terms of complexity levels, but can struggle with distinguishing texts with high complexity between each other.

RuBERT was fine-tuned as a regression model using a standard fine-tuning pipeline. The regression model was chosen as a means of modeling the relation

between the complexity levels and, and thus, produced the results in a way, where wrong predictions are relatively close to their real values.

The next part of the model is a data encoder, which outputs a vector of length 133 for each text. Vector values present a set of linguistic features.

The features are split into 10 general categories:

– basic metrics, traditionally used in the tasks of readability assessment;
– readability formulas, adapted for the Russian language;
– words of various part-of-speech classes;
– part-of-speech n-grams;
– general-language frequency characteristics of text lemmas;
– word-formation patterns;
– separate grammes;
– lexical and semantic features, multi-word expressions;
– syntactic features;
– cohesion features;

Data encodings and language model predictions are then passed to the final hybrid module. Thirteen approaches were tested and compared, using different models trained with or without additional language model predictions.

It was found that in all tests the usage of language model predictions provided a substantial improvement to the quality of predictions. Using a set of classification and regression metrics, it was found that the XGBoost model, trained on features and predictions, provides the best quality with accuracy, precision and F1 scores 0.78 or higher. This, surprisingly, holds true even for regression metrics, such as RMSE (with 0.06 error rate) and R2 (with 0.9479 coefficient of determination). The resulting model can be used as a hybrid model, feature based model or language modeling-based model.

4.4 Results and Discussion

4.4.1 Complexity Scores by Sub-style and (Non)domestic Status

Table 12, Table 13, and Table 14 below present the results of language complexity estimation for national law documents (ASSDs, LSSDs and JSSDs), and international law documents. Table 12 shows the results of the hybrid model, Table 13 shows the ruBERT predictions, and Table 14 shows the metrics-based complexity predictions.

Table 12 — Hybrid Model Predictions

Complexity	Administrative	Legislative	Justiciary	International
12	911	14002	26368	1522
11	13	516	31	46
10	12	256	37	49
9	1	5	0	0
8	1	17	0	0
7	0	2	0	0
6	0	4	0	0
4	0	5	0	0
2	0	3	0	0
0	0	3	0	0

The results show that the vast majority of all documents in all of large classes are rated by all models as maximally complex. For instance, if we take a closer look at the results of the hybrid model (see Table 12), complexity class "12" includes 97.1% of administrative sub-style documents, 94.5% of legislative sub-style documents, and 99.7% of justiciary sub-style documents of national law. In relation to all documents of international law the proportion of documents with complexity level of "12" is 94.1%.

The set of LSSDs is the most diverse in terms of complexity. Next is an explanation of how the models work on a complexity level of "0", which actually was not expected for this dataset. The hybrid model and the fine-tuned ruBERT model assign this complexity level to three documents, among which are, for example, Order of the RF Ministry of Education and Science "On the Coordinating Council of the

Table 13 — RuBERT Predictions

Complexity	Administrative	Legislative	Justiciary	International
12	917	14224	26385	1546
11	10	418	48	69
10	9	107	3	2
9	1	31	0	0
8	1	15	0	0
7	0	2	0	0
6	0	4	0	0
5	0	1	0	0
4	0	3	0	0
3	0	2	0	0
2	0	2	0	0
1	0	1	0	0
0	0	3	0	0

Table 14 — Metrics Predictions

Complexity	Administrative	Legislative	Justiciary	International
12	915	14638	26374	1607
11	2	4	0	0
10	0	71	0	0
9	0	1	0	0
8	15	18	3	2
7	0	4	0	0
6	2	3	0	0
5	0	3	0	0
4	4	66	59	8
2	0	2	0	0
0	0	3	0	0

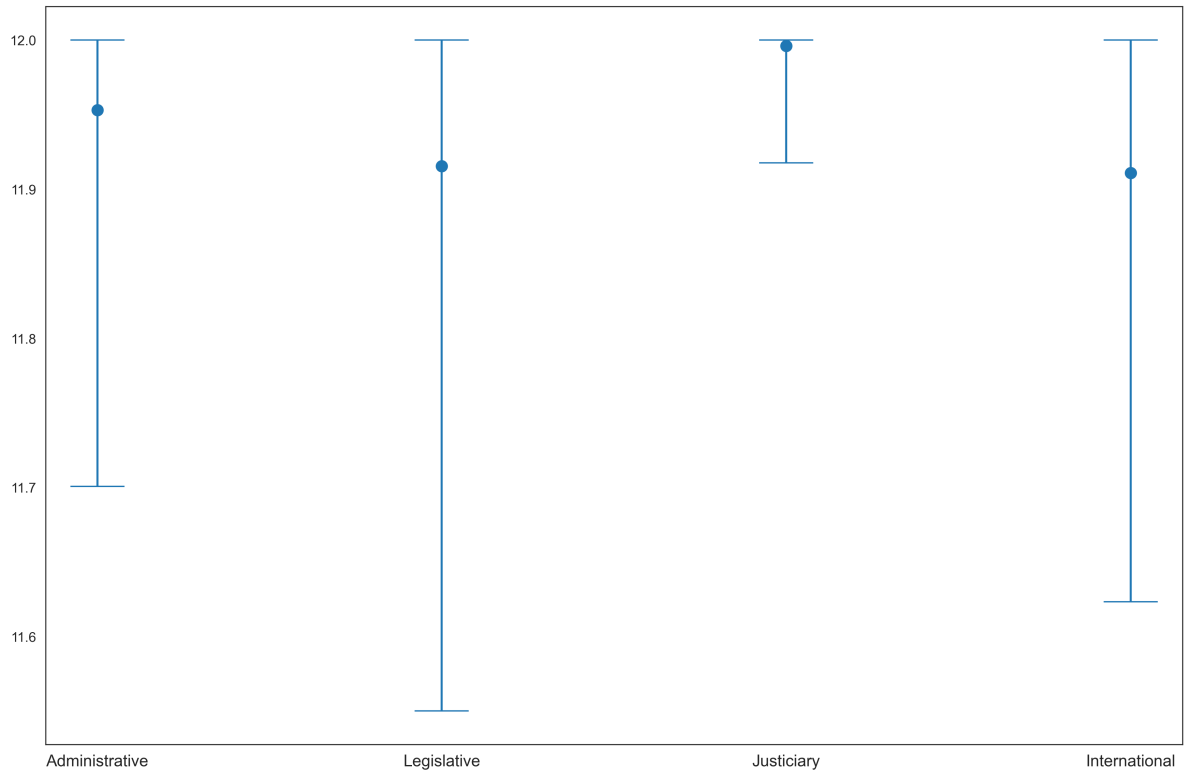


Figure 4.1 — Mean Values of Complexity (Hybrid Predictions)

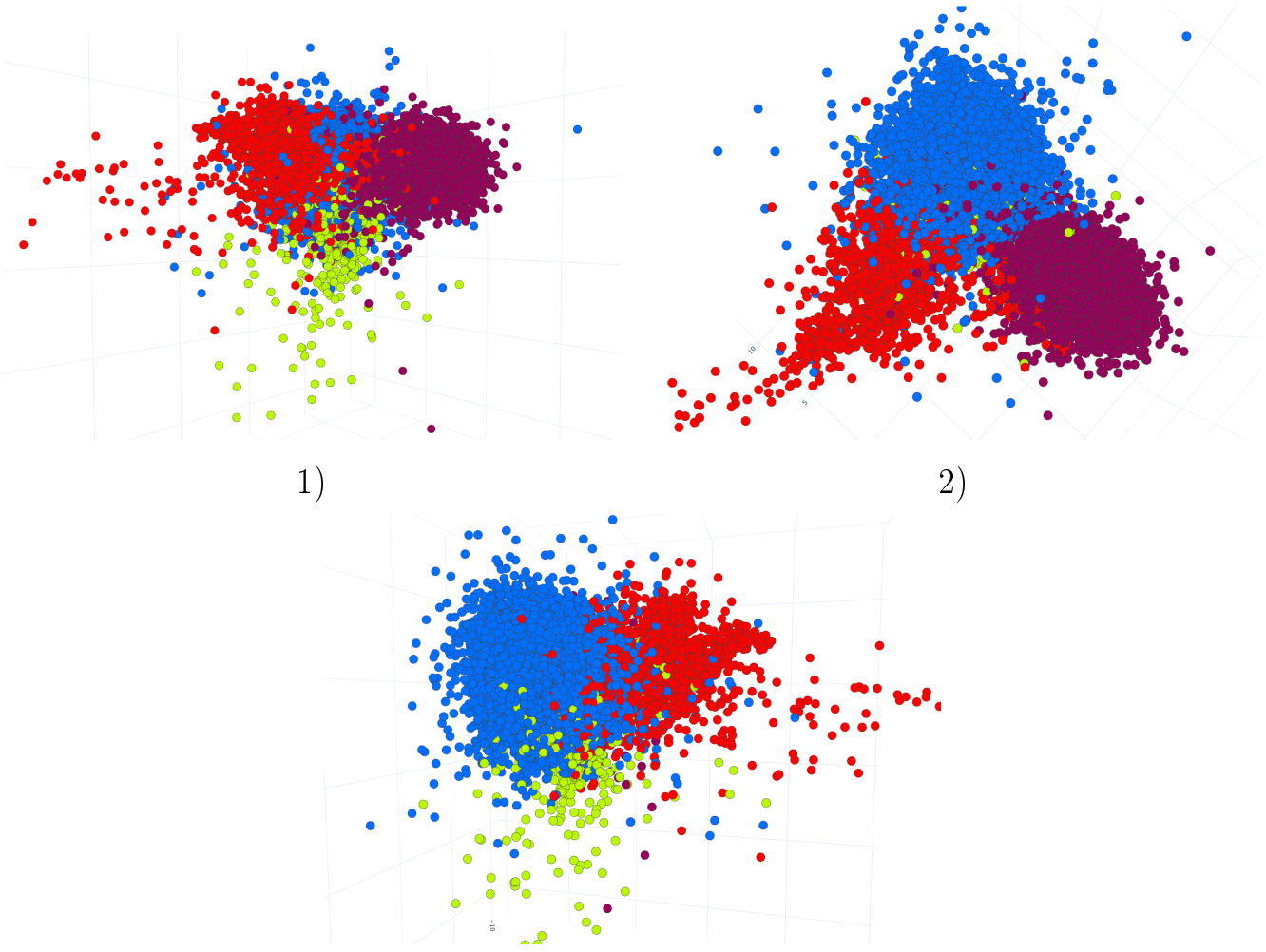
Ministry of Education and Science of the Russian Federation on the Modernization of Regional Preschool Education Systems". Thus, complexity level "0" is assigned to the documents whose subject matter relates to pre-school education. The metrics-based model assigns difficulty level "0" to other three documents, which are long sequences of short noun phrases with asyndetic coordination, see, for example, RF Government Decree of February 14, 2002 № 103 "On approval of the list of vital and essential medicines and medical devices for free acquisition by citizens permanently residing (working) in the territory of the zone of residence with the right to resettlement, in accordance with paragraph 19 of part one of Article 18 of the Law of the Russian Federation «On the social protection of citizens exposed to radiation due to the disaster at the Chernobyl nuclear power plant»".[5] At the same time, the Order № 103 contains many super-rare words (names of medicines), for example, "Allopurinol", "Trihexyphenidyl", "Carboplatin", and is defined by the fine-tuned ruBERT model and hybrid model as maximally complex text.

One-Way ANOVA on the complexity of each sub-style shows a significant difference between the means of different sub-styles with 278.4 F-value. Fig. 4.1 shows the mean values of complexity for each sub-style end status along with their standard deviations; complexity scores were obtained by the hybrid model.

The visualization confirms that the most complex documents in the studied dataset are JSSDs.

Linear Discriminant Analysis (LDA) was performed to reduce the dimensions of the feature vectors from 133 language parameters down to 3. Fig. 4.2 shows the visualization of sub-styles and statuses using the reduced vectors for each document.

Fig. 4.2, in particular, demonstrates that linguistic features well contrast between justiciary and legislative sub-style documents, while administrative sub-style texts are mixed with the texts of two other sub-style classes. In addition, it can also be argued that the values of linguistic metrics have successfully distinguished international and domestic legal documents.



1) 2)

3)

Figure 4.2 — Documents Comparison using LDA for Dimensionality Reduction (three projections)

For a more detailed comparison of documents by the status, the mean values of linguistic metrics were analyzed. To compare these values between national law documents and international ones a t-test was performed. It has been found that

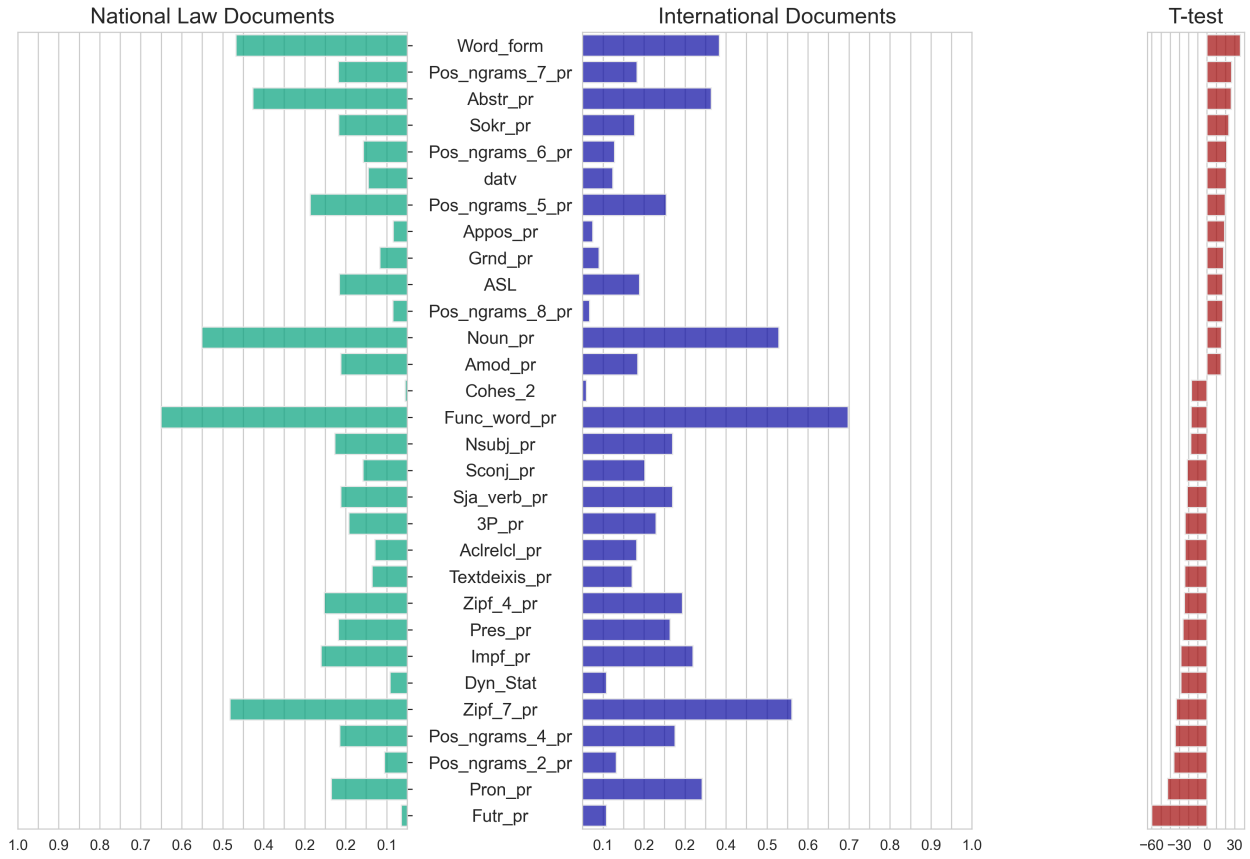


Figure 4.3 — Mean Values of Linguistic Metrics in Documents by Status

for Bonferroni adjusted p-values less than 0.05, null-hypothesis (equal mean values) can be rejected for 96 linguistic features, meaning there are significant differences between the mean values for these features. For p-values less than 0.01 and less than 0.001 the null hypothesis is rejected for 94 and 90 parameters respectively.

Fig. 4.3 shows the differences in mean values for national and international documents, normalized and sorted by the t-test statistic. For the purposes of plotting, only parameters which have t-test values greater than 15 are shown.

One can make some observations, according to which in domestic documents compared to international ones there are more derivative words, sequences of the type “noun + noun in the genitive case”, abstract words, graphic abbreviations, sequences of the type “noun + noun + noun”, appositive constructions, occurrences of adverbial participles. In addition, the sentences in the domestic documents are longer.

International documents as compared to domestic ones have more future tense verbs, occurrences of personal pronouns, sequences of the type “noun + finite verb”, sequences of the type “full adjective + noun”, and frequent lemmas (Zipf value =

7). Let us note also that (according to the dynamic/static formula) international documents are "more dynamic".

4.4.2 Complexity Scores by Genres

For each sub-style within the group of national law documents averages of specific categories of features were calculated, namely the "Syntactic", "Basic" and "Part-of-Speech" ones. Averages were calculated after the min-max normalization of each feature. Fig. 4.4, 4.5 and 4.6 present the averages and their respective standard deviations for each genre. The values of the averages on the visualizations are ranked by decreasing values of syntactic metrics. This solution will give a meaningful interpretation of the data obtained, since a very diverse distribution of domestic documents according to the complexity scores (see section 3.1 above for more information) was not obtained. Thus, generalizations based on syntactic features can be made, because they can be considered the most revealing in assessing text complexity.

Next are some comments on specific metrics. The list of syntactic features includes:

1. The features showing the structure of particular syntactic phrases (e.g. noun phrase, see the metric "Amod_p", i.e. the proportion of adjectival modifiers of a name; verb phrase, see the metric "Advmod_pr", i.e. the proportion of adverbial modifiers of a predicate);
2. The feature describing the occurrences of appositional modifiers ("Appos");
3. The features indicating the presence of coordinative series (meaning the feature "Cc" 'coordinating conjunction', and the feature "Conj" describing the number of conjuncts);
4. The features describing the occurrences of clausal modifiers of a noun (participles and participial clauses "Acl" separately from relative clauses "Acl:relcl"), adverbial clause modifiers, various clausal complements ("Ccomp", "Xcomp"); the units capable of attaching dependent clauses are counted separately ("Mark");
5. The feature describing occurrences of clauses with copula-like elements ("Cop");

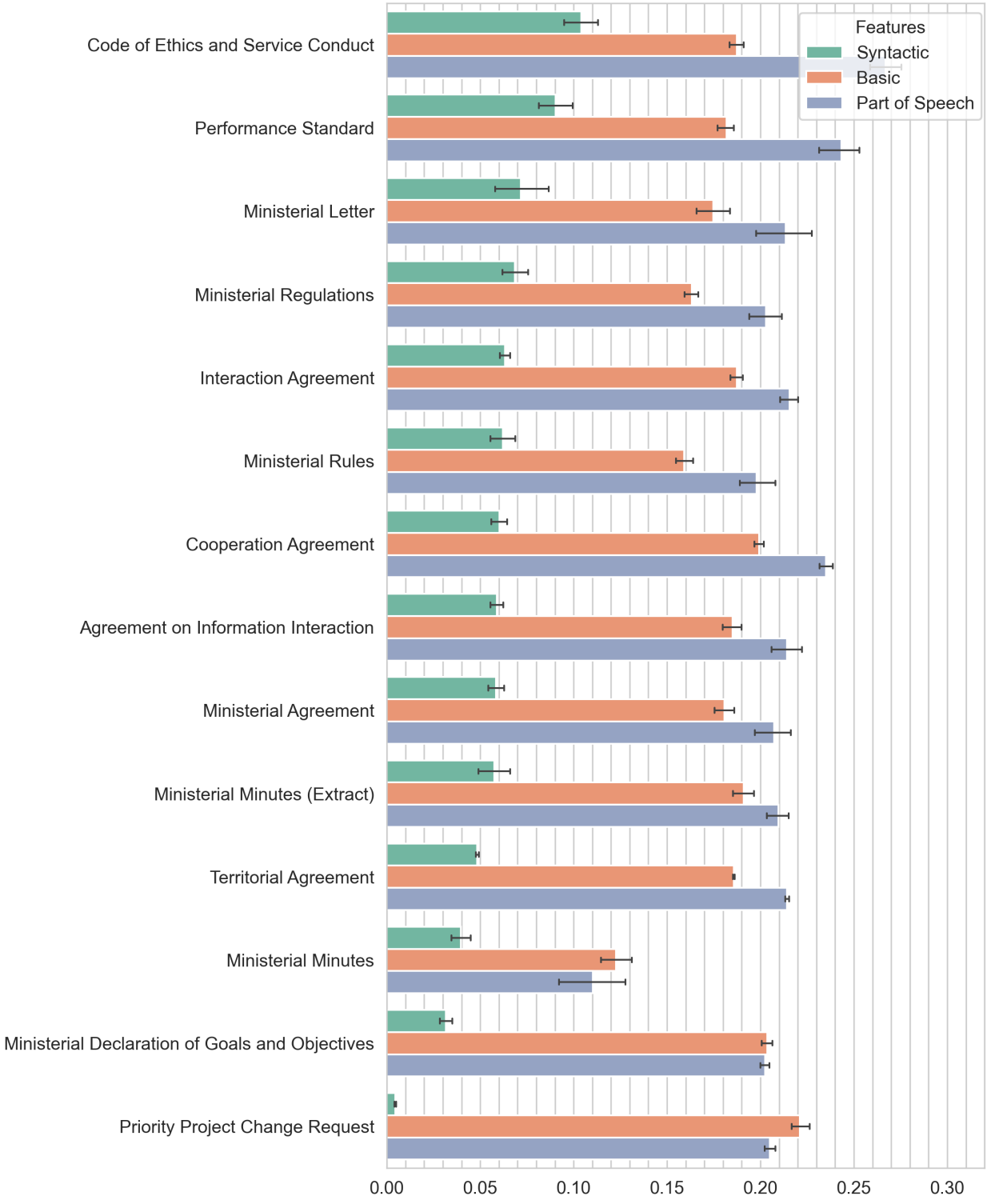


Figure 4.4 — Genres' Complexity within Administrative Sub-style

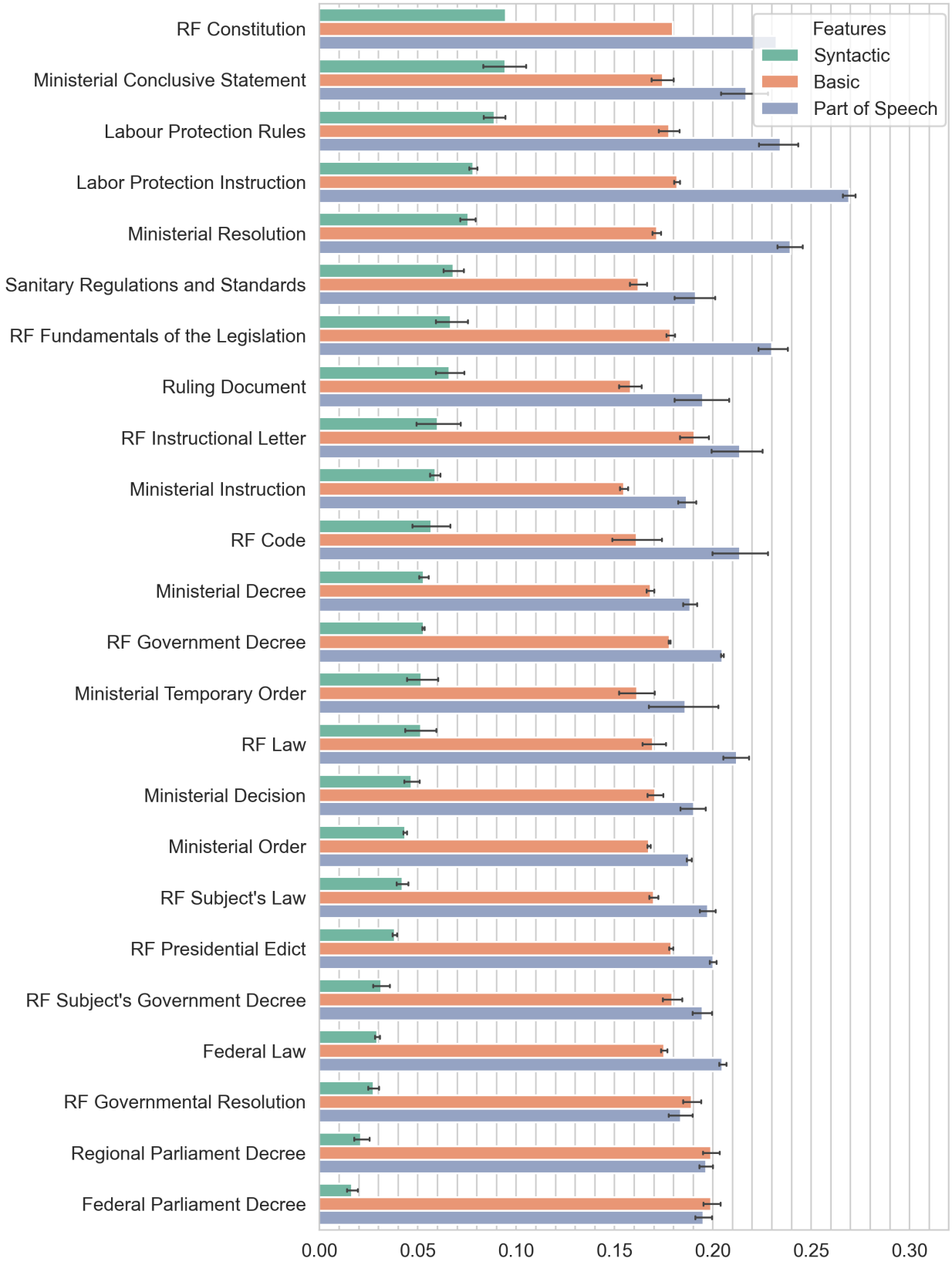


Figure 4.5 — Genres' Complexity within Legislative Sub-style

Figure 4.6 — Genres' Complexity within Justiciary Sub-style

6. The features that describe the occurrences of passive constructions ("Aux:pass", "Nsubj:pass", "Csubj:pass").

The possibilities of analyzing syntactic complexity are conditioned and limited by the parsing format. In this case, an important component of the complexity model is the consideration of features based on UDPipe markup [62]. Additionally, pymorphy2 was used for part-of-speech tagging and morphological annotation [111].

The main findings are as follows. Among the administrative sub-style documents, the Codes of Ethics and Service Conduct are the most syntactically complex ones. An example of a document of this genre is "Standard Code of Ethics and Official Conduct for State and Municipal Officials".[6]Legislative sub-style documents showed such a pattern: the most syntactically complex document surprisingly turned out to be the RF Constitution. Federal Parliament Decrees are the least syntactically complex (even though they have the highest complexity score according to basic metrics). As for justiciary sub-style documents, the most syntactically complex (with a noticeable break from other genres) are the decrees of the RF Constitutional Court.

In general, a comparison of the genre-based document groups (characterized in terms of the institutions that issued the particular texts) shows that in all three sets of sub-styles it is not the genre itself that may be decisive for the complexity score, but the issuing state authority or court. This can be clearly seen in the example of justiciary documents, in the set of which the decrees of the RF Supreme Arbitration Court and the decrees of the RF Constitutional Court are clearly opposed in syntactic complexity.

4.5 Chapter conclusions

This chapter explored a genre-diverse set of legal texts (43,804 documents, 118,768,028 words in total). The dataset includes international law documents (1,617 texts, 6,400,239 words) and national law documents. The latter are divided into three sub-styles, namely administrative sub-style (938 texts, 3,798,795 words), legislative sub-style (14,813 texts, 58,430,223 words) and justiciary sub-style (26,436 texts, 50,138,771 words). All domestic documents are categorized by genre and according to the institution that issued the document. A total of 68 legal genre classes (14 administrative, 24 legislative, and 30 justiciary ones) are identified.

All documents are assigned complexity levels ranging from “0” to “12”. In this chapter, the complexity predictions of the fine-tuned ruBERT model were analyzed, the predictions on 133 linguistic metrics, and the predictions of the hybrid model. The main results of the analysis of document complexity by sub-styles and genres are as follows.

The vast majority of all documents in all large classes are rated by all the models as maximally complex. Thus, the hybrid model assigns complexity class of “12” to 97.1% of administrative sub-style documents, 94.5% of legislative sub-style documents, and 99.7% of justiciary sub-style documents of national law. In relation to all documents of international law the proportion of documents with complexity level of “12” is 94.1%. The set of LSSDs is the most diverse in terms of complexity. On average, the most complex documents in the studied dataset are JSSDs.

Linguistic features well contrast between justiciary and legislative sub-style documents, while administrative sub-style texts are mixed with the texts of two other classes. The values of linguistic metrics have successfully distinguished international and domestic legal documents.

A more detailed comparison of documents by domestic/international status using t-test showed that there are significant differences between the mean values for 110 linguistic features. Specifically, in domestic documents compared to international ones there are more derivative words, sequences of the type “noun + noun in the genitive case”, abstract words, graphic abbreviations, sequences of the type “noun + noun + noun”, appositive constructions, occurrences of adverbial participles. In addition, the sentences in the domestic documents are longer. International documents as compared to domestic ones have more future tense verbs, occurrences of personal pronouns, sequences of the type “noun + finite verb”, sequences of the type “full adjective + noun”, and frequent lemmas (Zipf value = 7).

When comparing documents by genre, the average values of all syntactic metrics were interpreted . Averages were calculated after the min-max normalization of each feature. Among the administrative sub-style documents, the Codes of Ethics and Service Conduct are the most syntactically complex ones. The most syntactically complex legislativesub-style document surprisingly turned out to be the RF Constitution. Federal Parliament Decrees are the least syntactically complex (even though they have the highest complexity score according to basic metrics). As for justiciary sub-style documents, the most syntactically complex (with a noticeable break from other genres) are the decrees of the RF Constitutional Court.

In general, a comparison of the genre-based document groups (characterized in terms of the institutions that issued the particular texts) shows that in all three sets of sub-styles it is not the genre itself that may be decisive for the complexity score, but the issuing state authority or court [7]..

Chapter 5. Accessibility of legal texts

5.1 Introduction

A group of experts from St. Petersburg State University, commissioned by the Federal Taxation Service of Russia (hereinafter referred to as the FNS), assessed the level of accessibility of perception of written responses from tax authorities to requests from individuals and organizations.

The analysis was carried out on a collection of 2339 pairs of real questions and answers provided by Federal Taxation Service employees for research and assessment in 82 regions of Russia (from 2 to 48 question-answer pairs from each region). The assessment was carried out according to a methodology prepared by St. Petersburg State University based on the results of its own research in the framework of a number of scientific projects, including within the framework of the Research Institute of State Language Problems and devoted to the study of the language of official (legal) documents, and previously agreed upon by representatives of the Federal Taxation Service. The methodology involved an assessment based on automated machine analysis of the text of each response according to 12 criteria, for each of which a numerical assessment of communicative quality (accessibility to perception) was given and which were subsequently combined into an overall assessment for each material and an average assessment for each region.

Each of the evaluation criteria has a legal and linguistic justification based on what requirements for the language of official documents are directly provided for by the provisions of the Constitution (taking into account their interpretation by the Constitutional Court of the Russian Federation), current legislation and by-laws, and also obviously follow from these provisions. The fulfillment of these requirements must be assessed and carefully monitored.

The proposed criteria assume to a greater extent an analysis of the content of the document based on its communicative properties based on linguistic rather than content characteristics: the correctness of the answers was not assessed from the point of view of the correct explanation of tax legislation. Requirements for the form of the document, which are in the nature of independent legal requirements (details, signature by an authorized official, etc.) were not assessed, with the exception of

compliance with the rules of Russian spelling and punctuation, since errors made can significantly affect the certainty and understandability of the answer.

Each of the evaluation criteria provides for the use of software and tools, the use of which is not limited and does not require the consent of the owners of intellectual property rights, including any protected objects, the intellectual rights to which belong to St. Petersburg State University.

The entire set of automatic text analysis tools and its individual elements have not previously been used to solve the assigned tasks (assessing the communicative quality of tax authorities' responses to taxpayers' questions), therefore, after receiving the results of automatic processing and automatic evaluation of texts, selective manual control and evaluation was carried out to control the results obtained adequacy of the data obtained. In addition, the results of automatic assessment according to different criteria were summarized and compared.

Some criteria were not provided with effective verification tools in the context of these specific tasks. In particular, the search for colloquial vocabulary yielded practically no meaningful results, but analysis of the data obtained gives reason to believe that this was largely due to the lack of a dictionary of colloquial vocabulary suitable specifically for these tasks. The search for low-frequency words—words that are not found in the frequency dictionary based on the National Corpus of the Russian Language (an extensive representative collection of Russian texts designed to present the state of the modern Russian language in all its diversity) are not found or are found very rarely—provided not very effective results.

At the same time, for a number of criteria, automatic analysis gave indicative results. First of all, this concerns the overall assessment of text understandability using 19 metrics developed and used in modern linguistics. The analysis made it possible to divide all analyzed texts into groups, comparing them with each other according to characteristics that affect the ease and accessibility of perception, and texts that received low scores on one metric, as a rule, received low scores on other metrics, which confirms their overall effectiveness.

Effective results were obtained based on the "legal" nature of the response texts - the presence of special legal terms, verbatim quotes from regulations and paraphrased fragments of regulations. All these criteria significantly complicate the perception of the text by non-professional addressees and at the same time demonstrate low communicative efficiency: if the answer largely consists of what is contained in regulations, then this is most often not what applicants want to see.

It can be concluded that the proposed tools, in general, quite successfully solve the task of assessing those documents that were presented for analysis, but for large-scale systemic use they require adaptation to the assigned tasks based on the results of testing their use.

5.2 Evaluation criteria

5.2.1 Basic criteria

Spelling and grammar From a legal point of view, violation of spelling and grammatical rules creates a risk of uncertainty in the content and affects the perception of the received answer. In addition, a violation of spelling rules can create uncertainty as to which word should actually be contained in the text. The analysis showed that 30% of all answers (710) did not have a single spelling or grammatical error.

Punctuation Violation of the rules of punctuation further creates uncertainty of the content, since the lack of correctly placed punctuation marks allows for different interpretations of the syntactic structure of the sentence. The analysis showed that 16% of all answers did not have a single error in the placement of commas. In the remaining answers, errors are isolated. Moreover, almost all analyzed answers received very high scores for this criterion - from 96 to 100.

Figures of speech and idioms According to the criterion of the presence of figures of speech and idioms in the answer texts, an assessment was made on a general scale from 0 to 100, where 100 received an answer text in which not a single metaphor or idiom was found, and 0 - an answer text in which the maximum relative number of words and expressions with the specified characteristics is found. Each answer was scored on this scale depending on its position within the minimum and maximum values of the number of words being assessed.

To check for the presence of idioms (stable non-single word sequences of words, the meaning of which is not directly derived from the meaning of the words included in the sequence), the MMFLD library (https://github.com/laihuiyuan/MMFLD), based on the T5 architecture, was used. Sequences of the form were fed to the input

of the model: 'Which figure of speech does this text contain? (A) Literal. (B) Idiom. | Text: Suggestion of response text'. Thus, the model determined the presence of idioms in individual sentences of the response text. Based on this limitation, the assessment was based on the number of sentences in the text.

As a result of the analysis, not a single example of the use of metaphors was found in the texts under consideration. The presence of figures of speech and idioms in the texts of the answers is sporadic. The vast majority of analyzed answers for this criterion received the maximum score.

Colloquial vocabulary According to the criterion of the presence of colloquial vocabulary in the response texts, an assessment was made on a general scale from 0 to 100, where 100 received a response text in which not a single word related to the colloquial style of speech was found, and 0 - the response text, in which the maximum relative number of words and expressions with the specified characteristics is found. Each answer was scored on this scale depending on its position within the minimum and maximum values of the number of words being assessed.

Low-frequency words According to the criterion of the presence of low-frequency (rare, little-used) words in the response texts, an assessment was made on a general scale from 0 to 100, where 100 received a response text in which not a single word related to low-frequency words was found, and 0 – the text of the response in which the maximum relative number of words and expressions with the specified characteristics was detected. Each answer was scored on this scale depending on its position within the minimum and maximum values of the number of words being assessed. Low-frequency words were selected from the list of the "New Frequency Dictionary of Russian Vocabulary" with a Zipf measure of less than 3. The Zipf measure is based on a logarithmic transformation of the ipm frequency value (relative frequency, the number of occurrences per million words of the text collection on the basis of which the frequency list of words was compiled, ranked in descending order of frequency of occurrence) and allows to distribute all words present in a certain frequency list into ranges (and thus separate high-frequency, mid-frequency and low-frequency units, and then estimate the number of low-frequency units).

Sentiment (emotional characteristic of the text) Assessing sentiment is not directly related to the consideration of the accessibility of the answer for perception and understanding, however, it allows us to formalize the concept of "general impression of the text." Sentiment analysis is the process of analyzing text and determining its corresponding emotional tone. It can be used to determine

the overall sentiment of a piece of text, such as whether the sentiment is positive, negative, or neutral. As a result of the model's work, each response text was assigned a ranked sentiment index, in which a value of 100 was assigned to a neutral text, 50 to a text with a positive sentiment, and 0 to a text with a negative sentiment. Thus, texts with a neutral tone (non-evaluative and unemotional, that is, meeting the requirements for texts in an official business style) received the highest score. Sentiment assessment was carried out using the pre-trained Rubert model. The analysis showed that all the texts under consideration have a neutral tone. This fully corresponds to the expectations of experts, since the responses of the tax authorities should have had a neutral tone, and any deviations from this would be considered a defect.

Formal criteria According to the criterion of the level of compliance of the response with formal rules, an assessment was made on a general scale from 0 to 100, where a value of 0 was assigned to a response in which the requirements for the form of application were violated and there was no necessary indication of the informational nature of the response, 50 was assigned to a response where If there was one of the two specified violations, 100 was assigned to a response in which the specified violations were not identified.

The criterion of compliance with the form of address was checked by a simple text comparison, the criterion of indicating the informational nature of the answer was checked using a coding language model. Using the basic pre-trained Rubert-base-cased model, sentence encodings of texts and text encodings were obtained. indicating the explanatory nature of the answer. If a cosine similarity value of encodings was found to be greater than 0.75, the response was considered to contain a warning message (a value of 0.75 allows warnings to be detected in the most free form).

For the analysis, two requirements contained in departmental acts of the Federal Taxation Service of Russia were taken into account - an indication of the informational nature of the response and compliance with the form of contacting the applicant. If necessary, new parameters can be added to the evaluation mechanism. The results obtained demonstrate that both requirements are met only in a small part of the analyzed responses. The vast majority of responses did not indicate their informational nature.

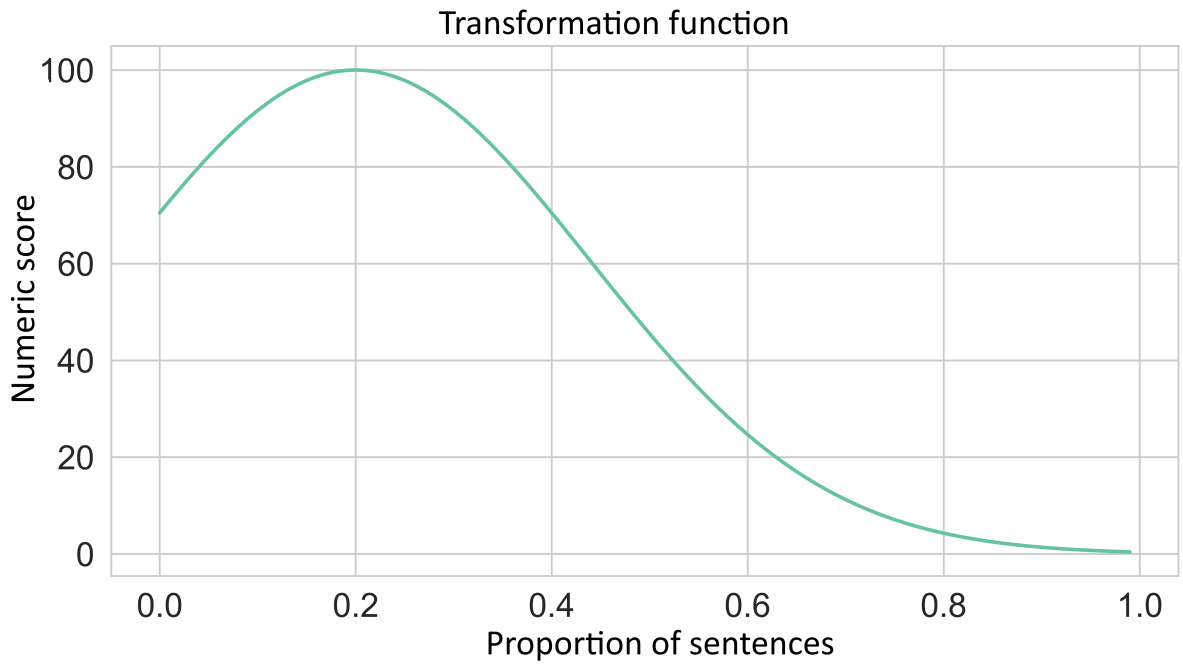


Figure 5.1 — Function for converting proportional values to a grading system from 0 to 100

5.2.2 Legal terminology

Based on the criterion of the presence of legal terminology in the texts of answers, an assessment was made with reduction to a normal distribution with a shift. In the texts of the answers, words and non-word terms were identified that were classified as words of legal terminology in the extensive legal dictionaries of A. B. Borisov and V. N. Dodonov. The list of received terms was refined using manual evaluation - words that, in the opinion of experts, had a clear "non-legal" meaning were removed. Next, the relative number of legal terms was calculated relative to all words of the text. From it, using the formula $100e^{-1.4((x-0.2)/0.4)^2}$, where x is the relative number of legal terms. Fig. 5.1 shows a visual representation of the formula.

This formula (a special case of normal distribution) allows to more smoothly set the optimal value for the number of legal terms. A formula with these parameters gives a maximum score of 100 for answers in which the share of legal terms is 20% of the total number of words. In the absence of legal terms, based on the formula, the answer receives a score of 70. As the proportion of legal vocabulary increases, the score approaches 0 according to the normal distribution.

In the test data - for 860 documents (37% of the total set of answer texts), the share of legal terms does not exceed 20% of all words in the answer text.

The analysis showed that in most responses from tax authorities, special legal terminology is used in optimal proportion to the volume of the responses themselves, which does not make it difficult to understand their content.

5.2.3 Matching question and answer

When assessing this criterion, attention was paid not only to the text of the response, but also to the question itself contained in the taxpayer's appeal. The test of the criterion was carried out using three models - two models of answering questions and a model of implication between sentences.

Models for answering questions (mdeberta-v3, xlm-roberta) require as input the text of the question and the text of the context, from which the model must select a short and succinct answer to the question. The presentation of the results is the same for the two models - the models receive the answer from the context and the degree of confidence in the choice. The latter is used as one of the evaluation parameters. For the final assessment, the maximum confidence value of the two models is taken.

The sentence sequencing model is part of one of the basic language modeling tasks that allows for obtaining the probability of the consequence of sentences in the text. In this case, it is believed that this model can provide the simplest assessment of the presence of general context in the texts of questions and answers.

The approach based on question answering models has serious disadvantages related to the ability to find concise answers to complex questions. For general language data, it is often enough to indicate a few words from the text that give a short answer to the question, which is not always true for answers to complex legal questions. In this regard, this criterion parameter was assigned a reduced weight of 0.3. The sentence consequence model was given a weight of 0.7. Fig. 5.2 shows the distribution of question and answer consistency scores.

It is important to note that a low criterion value could potentially be assigned to a detailed answer to a long and complex question. This factor is due to the

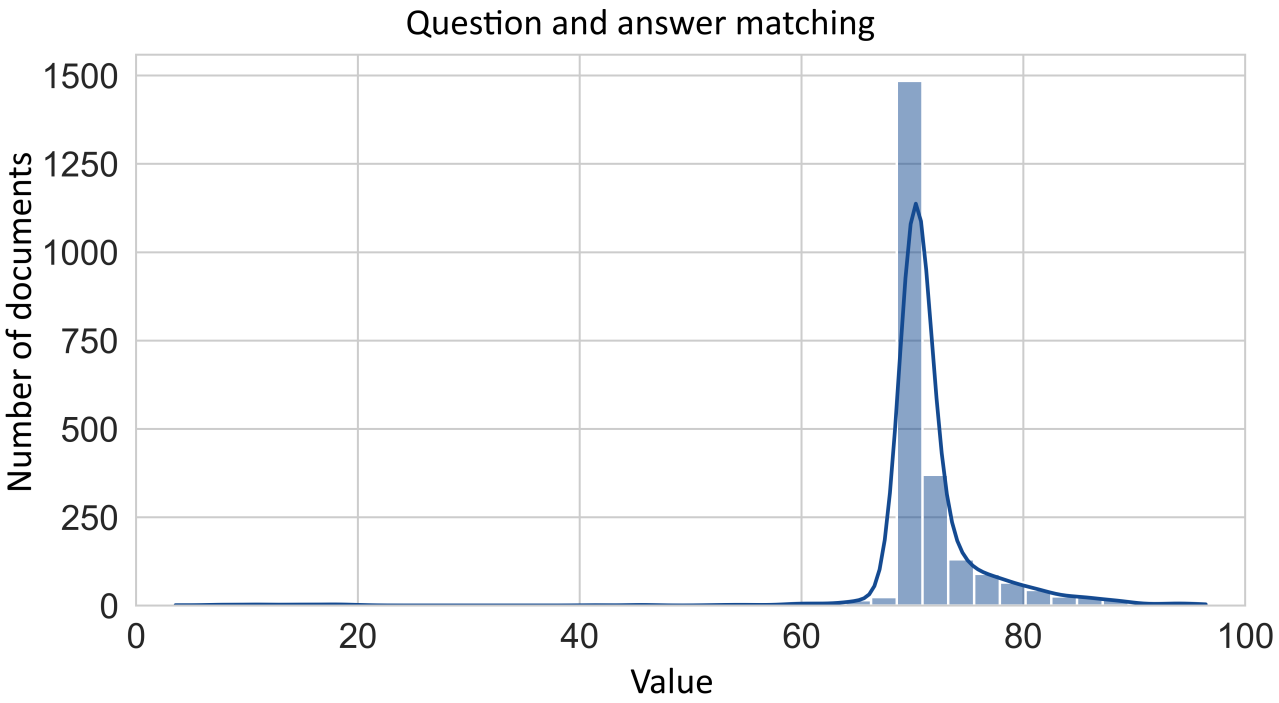


Figure 5.2 — Distribution of Questions and Answers Matching Scores

specifics of the work of neural network algorithms, which have a limited context window, and the specifics of the data.

The analysis demonstrated that, in most cases, the tax authorities' responses corresponded to the applicants' questions. The decrease in score was influenced by factors such as: excessive length of the answer; the presence in the response of information not directly related to the applicant's question; quoting regulations without the necessary explanations, lack of a concentrated conclusion in the answer, etc.

5.2.4 Paraphrases and quotations

To assess the presence of paraphrases of passages of legal documents in the texts, encodings obtained by the Rubert-base-cased model were used. Each sentence of the answer text was associated with a corresponding numerical vector. The encoding vectors of document sentences were created in a similar way:

– Constitution of the Russian Federation;
– Tax Code of the Russian Federation (parts one and two);

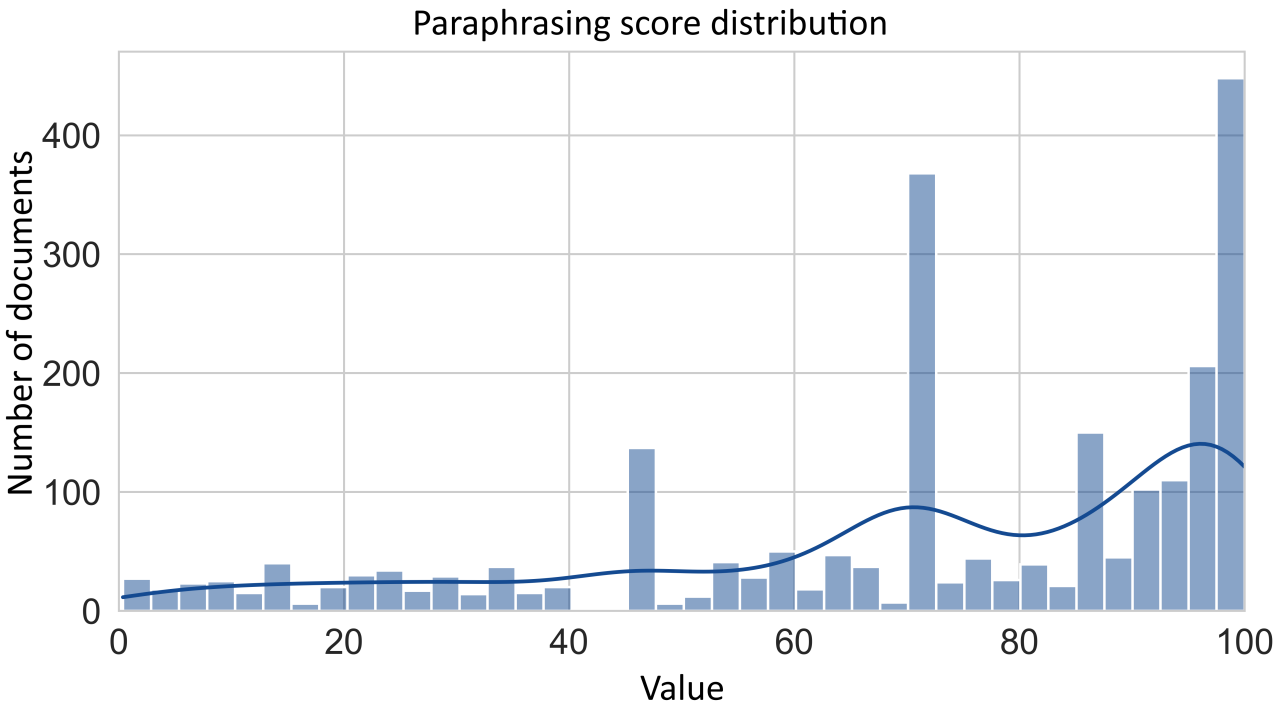


Figure 5.3 — Distribution of ratings for the presence of paraphrases

– Chapter 15 of the Code of the Russian Federation on Administrative Offenses.

Next, for each sentence of the response text, the maximum value of similarity (in the described case, cosine similarity) with the texts of the cited documents was obtained. Then the proportion of sentences with a similarity level greater than 0.85 but less than 0.95 was calculated. Finally, using a ranking approach based on a normal distribution with a shift, numerical scores for the level of paraphrasing of each response text were obtained. Thus, a score of 100 was given to responses in which 20% of the text consisted of paraphrased versions of sentences from the documents presented in the list. Completely original answers have a score of 70, answers that are entirely paraphrased have a score of close to 0. The resulting distribution of ratings for the presence of paraphrases is presented in Fig. 5.3. A total of 17 documents from the entire data set are identified as paraphrased fragments of legal acts.

The analysis showed that the responses of tax authorities very often contain excerpts from regulations. Moreover, if we take into account the assessment according to the following criterion - “Quotations”, most of these links have the form of indirect citations, and are presented in a relatively paraphrased version, adapted for the purpose of answering the applicant’s question. This significantly increases the

level of perception of the answer. In addition, it is important how much the answer is "loaded" with such references to regulations. The graphs presented above demonstrate that in most cases the answer consists of 0-20% of such paraphrased quotes. A significant part of the answers consists of 20-40% paraphrased quotes. It is also noteworthy that there are answers that almost entirely consist of similar quotes, which has a very negative impact on the level of their difficulty in understanding.

Direct quotations were determined similarly to the paraphrasing model. Vectors were created representing answer texts, document tests, and the degrees of similarity of sentences were obtained. Next, the proportion of sentences in the answer text for which the maximum degree of similarity exceeded 0.95 was determined.

A score of 0 was given to responses consisting entirely of direct quotes; a score of 100 was given to answers in which no direct quotes were found. Each answer was scored on this scale depending on its position within the minimum and maximum values.

Using a language model instead of a direct comparison has allowed for identifying direct quotations with minor changes (spelling errors, minimal insertions, etc.).

The resulting distribution of citation presence scores is presented in Fig. 5.4. More specifically, 1694 answers (72%) from the analyzed set do not contain direct quotations from legal acts. No response texts consisting entirely of quotes were found.

The results of the analysis show that in the responses of the tax authorities, direct citation of regulations are used, but this is done in acceptable quantities that do not have a serious impact on the perception of the content of the responses. In most answers, there are no direct citations of regulations at all. It is worth considering the prevalence of the use of paraphrased quotes described above.

5.2.5 Complexity

In itself, the complexity and accessibility of the response text is a requirement for the language of the tax authorities' responses, based on the fact that the clarity of the text ensures the taxpayer's right to receive complete and accessible information regarding his rights and obligations. Some of the other criteria described here are indirectly aimed at assessing the understandability of the text for perception, but

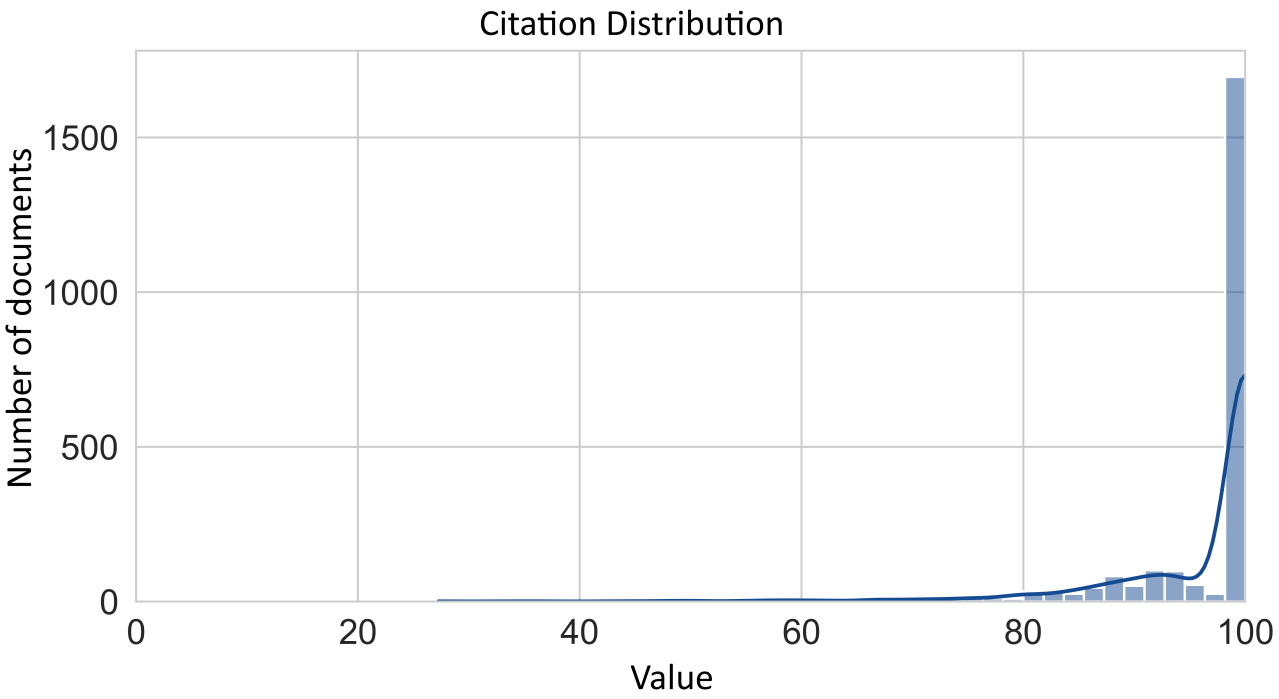


Figure 5.4 — Distribution of citation availability scores

in addition to this, an integrative parameter for assessing the understandability of the text of the answers should also be provided.

For each of the response texts, the value of each of the following 19 metrics was calculated (arranged in descending order of importance):

- FRE_GL adapted Flesch-Kincaid formula;
- SMOG adapted SMOG formula;
- ARI adapted formula for calculating the automated readability index;
- Nouns_pr noun vocabulary index;
- Inan_pr share of inanimate nouns;
- Adjif_pr share of full adjectives;
- ACW average length of a word form in letters;
- Gen_pr share of word forms in the genitive case;
- CLI index Colman-Liaw;
- word_long_pr proportion of long words (4 or more syllables);
- Adj_pr adjectivity index;
- ASS average sentence length in syllables;
- Prtf_pr share of full participles;
- DCI index Dale-Chale;
- ASW average length of a word form in syllables;

- Abbr_pr share of abbreviations;
- TTR_word simple TTR (word forms);
- N number of numeric characters;
- Prts_pr share of short participles.

"Complexity" is understood as an objective parameter that can be assessed in natural language texts. The "complexity" of a text, in turn, has a direct impact on its understandability for a particular reader. Complexity metrics are used to assess complexity. The choice of metrics is justified by linguistic experience in assessing text complexity.

The above list of metrics includes, first of all, automated readability indices (ease/difficulty of text to read), namely FRE_GL, SMOG, ARI, CLI, DCI. When calculating readability formulas, a simple logic is applied, according to which long (syntactically complex) sentences are more difficult to read and interpret than short ones; long words (in a particular case, words longer than four syllables) are more difficult than short ones, etc. Despite the simplicity of the approach, readability formulas have shown their effectiveness in assessing the readability of text in dozens of natural languages. It is worth adding that in the presented scheme for assessing the difficulty of reading (and, accordingly, the understandability of the text for the reader), only readability formulas adapted to Russian texts are used.

In addition, the list of metrics includes Inan_pr - the proportion of inanimate nouns. This metric, along with Nouns_pr (proportion of nominal vocabulary), is designed to capture the introduction of concepts denoted by nouns into texts (this parameter is sometimes called "lexical density"). According to general logic, the more concepts there are in a text, the more complex it is (this fact has been confirmed by a number of previous studies). It is worth adding that many inanimate nouns in the texts of legal documents are most likely abstract (non-objective). Experimental linguistic studies show that abstract vocabulary is more difficult to interpret than object nouns. The metric "proportion of complete adjectives "Adjif_pr"" is also intended to assess both syntactic and conceptual complexity. Thus, according to the "adjective plus noun" model, a number of terms and term-like combinations are formed (cf. "single agricultural tax"), which, due to its rarity and specialization, has not yet been included in general dictionaries of legal terms.

The metrics "proportion of full participles (Prtf_pr)" and "proportion of short participles (Prts_pr)" are intended to describe the syntactic complexity of texts in responses to requests. Participial phrases are not considered sentences, but linguists

call such phrases "participial clauses." Participial clauses, like sentences, are predications (as are, for example, simple sentences with finite forms of verbs, cf. "except for cases provided for by law" vs "except for cases provided for by law"). The presence of participial clauses in the general case makes the syntactic structure of the text more complex, which is confirmed by a number of studies.

Further, the list of metrics includes the proportion of word forms in the genitive case. This parameter indirectly describes the syntactic complexity of the text, including the presence of chains of nouns in the genitive case (cf. "failure to confirm the possibility of fulfilling the customer's request", "impossibility of applying the assumption of continuity of the enterprise"), as well as the occurrence in the texts of non-literal terms and term-like combinations formed according to the model "noun in the nominative case + noun in the genitive case" (cf. "summa naloga" (tax amount), "forma deklaracii" (declaration form), "kabinet nalogoplatel'shchika" (taxpayer's account)).

The "TTR" metric was introduced to measure the lexical diversity of a text (simply put, the higher the metric values, the more different words there are in the text, and the less words are repeated). It is generally believed that texts with low lexical diversity are easier to read.

The "Abbr_pr" metric (proportion of abbreviations) is used because in special texts (including legal texts on different topics) there are a significant number of abbreviations that are familiar to professionals, but unfamiliar to "ordinary" native speakers. These abbreviations are often not explained in the text, but their interpretation is obviously difficult.

The "N" metric (number of numeric characters) is designed to record the presence of numbered lists in the text, which in legal texts tend to be cumbersome and extensive. In addition, the metric indirectly reflects the presence of various references to the provisions of legal acts (cf. "clarifications received from the Ministry of Finance of the Russian Federation on this issue (letter of the Ministry of Finance of Russia dated November 12, 2004 No. 03-03-02-02/10)") and occurrences of quantitative expressions in texts.

Finally, the list of metrics contains easily interpretable indicators (average word form length in letters, proportion of long words, average sentence length in syllables), the rationale for their use is outlined in the passage above on readability formulas.

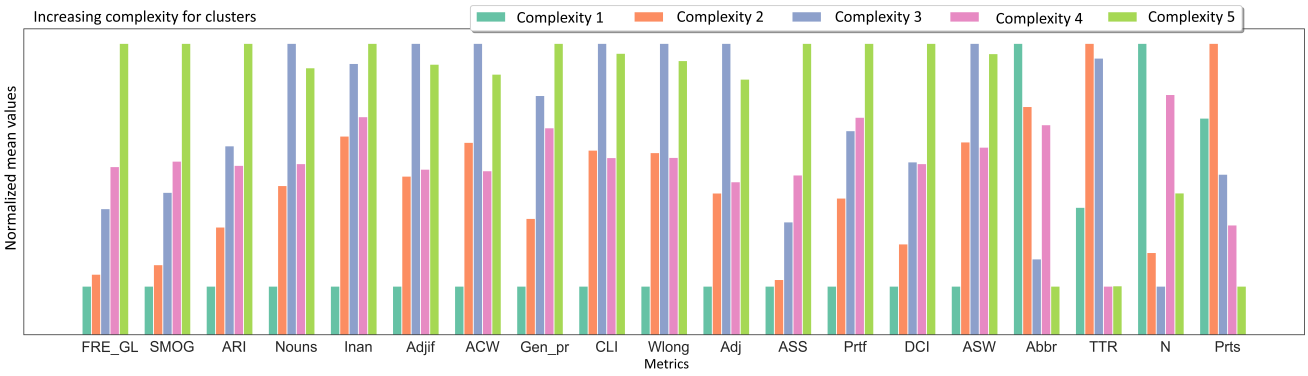


Figure 5.5 — The increasing nature of the complexity of clusters as seen in the metrics

Clustering was used to evaluate response responses. After calculating the values of these metrics, the KMeans clustering model was used as the initially proposed HDBSCAN model showed unsatisfactory results with all combinations of hyperparameters, despite its high efficiency in text analysis tasks [6], the responses were distributed into 5 clusters.

Fig. 5.5 shows the increasing nature of the complexity of clusters. Based on the visualization, it can be seen that the most important criteria (the first ones on the list) have a direct correlation with the complexity index. This rule is not fulfilled for the latter criteria, which indicates their insignificant influence on the complexity assessment in the data set under consideration.

A brief interpretation of the values of the three metrics that make the greatest contribution to assessing the complexity of answer texts.

- **FRE_GL** (adapted Flesch-Kincaid readability formula). The variables in the formula are ASL (average sentence length in words) and ASW (average length of word forms in syllables). High metric values generally mean that texts contain many long (and therefore syntactically complex) sentences and many long (4 or more syllables) words. The value of this metric, when looking at the entire analyzed data set, varies widely from 1.56 to 73.64;
- **SMOG** (adapted Simple Measure of Gobbledygook readability formula). The variables in the formula are the values of "number of sentences" and "number of long words". The SMOG value is calculated as the ratio of the number of long (respectively, complex) words to the number of sentences. Thus, the metric is able to find sentences with a large number of long words;

– **ARI**(adapted formula for calculating the automated readability index). The variables in the formula are the number of characters, the number of words, and the number of sentences of the text. More precisely, the index takes into account the ratio of the number of characters to the number of words and the ratio of the number of words to the number of sentences. This means that the lengths of words and sentences are again considered, however, when estimating the length of words, not syllables, but signs are used (which makes it possible to estimate the length of digital and alphanumeric complexes like 03-04-07/102199).

The complexity score was calculated inversely proportional to the complexity cluster index - answers that fell into the highest complexity cluster received a score of 0, and those in the low complexity cluster received a score of 100.

The cluster of the greatest complexity (with a score of "0") included 299 answer texts (13%), the cluster of increased complexity (with a score of "25") - 613 answer texts (26%), the cluster of average complexity (with a score of " 50") – 446 answer texts (19%), in the low complexity cluster (with a score of "75") – 635 answer texts (27%), in the low complexity cluster (with a score of "100") – 345 answer texts (15%).

Interpretation of the meaning of metrics for specific response texts made it possible to establish the following main factors influencing the understandability of response texts.

1. Formation of response texts from non-paraphrased fragments of regulatory legal acts;
2. The use of template expressions (often long and incomprehensible), which can be replaced by shorter and simpler expressions without loss of meaning (cf. "via telecommunication channels" vs. "by email");
3. The use of long enumerative ("composed") series, moreover, not designed in the form of numbered lists that can somehow facilitate understanding;
4. The use of lexical (verbal) repetitions, which increase the length of sentences in words and can be removed from the text without loss of meaning.

The analysis demonstrated the high efficiency of the mechanisms used. At the same time, the generalization of assessments by region deserves special attention. This mechanism can not only identify individual "incomprehensible" answers, but also use a large amount of data to determine "problem" areas - these can be any elements involved - from regional departments to specific specialists who prepared the answer (subject to availability in the materials submitted for analysis this data).

5.2.6 Combined score

To obtain a final score combining the criteria, in accordance with their importance and quality of assessment, each criterion was given an appropriate weight.

The understandability criterion received a weight of 0.2, the criterion for the presence of legal terms - 0.1, the paraphrasing and citation criteria received a weight of 0.15, and the remaining criteria received a weight of 0.05.

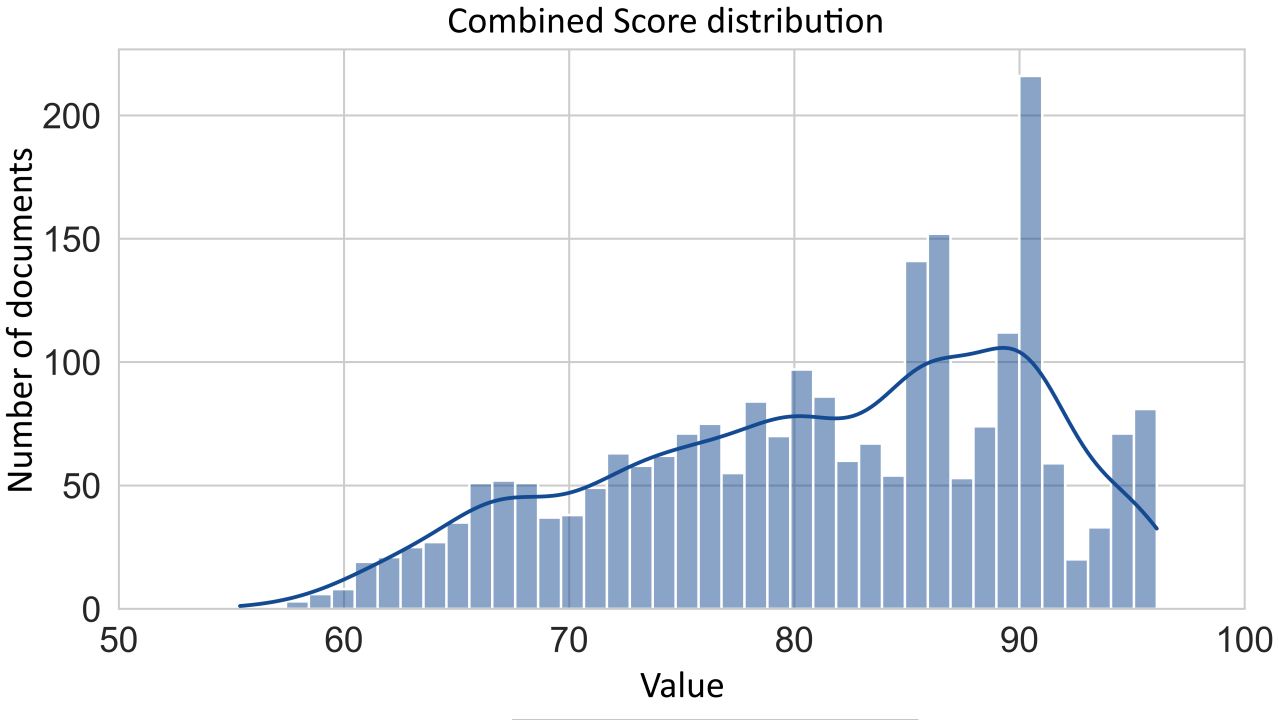


Figure 5.6 — Distribution of final combined scores

Figure 5.6 shows the final distribution of combined response scores. Thus, the analyzed responses from tax authorities have a high communicative quality (score from 80 to 100) in 58% of responses. Average communicative quality of answers (indicator from 60% to 80%) is observed for 41%. Less than 1% of responses have acceptable communication quality (indicator from 40% to 60%).

Conclusion

The analysis of legal texts is a relevant and important area for the country. The development of information technologies in the field of legal applications (LegalTech) is a promising subject, important both in the scientific context and in the practical one.

Based on the conducted research, the following conclusions were made regarding the development, relevance and applicability of various methodological approaches to the processing of legal documents.

The presented hybrid approach, combining traditional methods of text data analysis and modern neural network approaches in language modeling, allows to achieve a high quality assessment of complexity. Traditional methods of statistical analysis and machine learning make it possible to identify the language characteristics that have the greatest impact on the final complexity value. Thus, the methodology combines the accuracy of predictions of neural network approaches and the interpretability of classical approaches.

The methods and software solutions presented in this work have been used in research projects devoted to the analysis of the complexity of various types of legal documents. Certain methods have shown high efficiency in the practical task of analyzing texts of answers to questions in the legal and economic spheres.

The presented methodology takes into account various features of legal documents, such as the specificity of the language and structure of documents, inadequate amount and variety of publicly available data, and a wide variety of texts of various genres. This feature allows for high adaptability the process as a whole and individual analysis methods for tasks that have similar data characteristics.

The purpose of the work was to develop and test methodological and instrumental tools for the intelligent processing of legal texts and algorithmic support for the process of determining the accessibility of their perception. To achieve the goal, the following tasks were solved:

- The current state of legal and linguistic research in the field of analysis of legal documents has been studied, current problems have been identified and methods for solving them have been proposed.
- Methodological approaches have been developed for collection, processing and semantic analysis of the Russian legal language.

- A methodology has been developed for statistical assessment of the frequency characteristics of legal language.
- The linguistic characteristics of legal documents that best describe them in the context of complexity and accessibility have been identified and selected.
- A program architecture has been developed for analyzing the complexity of legal documents based on methods of hybrid use of language models.
- A comparative analysis of the complexity of documents of various substyles and genres was carried out using a hybrid complexity assessment model.
- A practical analysis of the complexity and accessibility of legal texts was carried out using the presented models and methods.

The author expresses gratitude and appreciation to the scientific supervisor Ivan Stanislavovich Blekanov for his support, assistance in creating, discussing the results and scientific guidance. The author also thanks Olga Vladimirovna Blinova for leading the project "Understanding official Russian: the legal and linguistic issues", for expert linguistic assessment of the quality of methods and co-authorship in key articles that became the basis of this dissertation work. The author thanks the head of the Research Institute for Problems of the Official Language, Sergei Aleksandrovich Belov, for consultations and expert legal assessments. Bodrunova Svetlana Sergeevna for joint work in projects devoted to semantic analysis of text data. The author thanks everyone who made this work possible.

Bibliography

1. *Certificate of state registration for a program.* A program for identifying echo chambers in discussions on social media platforms based on an analysis of the polarization of user opinions (SNAOpinionPolariz) [Text] / I. S. Blekanov, N. A. Tarasov, S. S. Bodrunova ; Rospatent. — № 2023685490 ; заявл. 27.11.2023 (Рос. Федерация).
2. *Certificate of state registration for a program.* A program for automatic summarization of user messages in social network discussions (SNAPostSummarizer) [Text] / I. S. Blekanov, N. A. Tarasov ; Rospatent. — № 2021680151 ; заявл. 21.11.2021 (Рос. Федерация).
3. *Certificate of state registration for a program.* A program for automatic detection of hidden topics in user discussions on social networks (SNATopicDetector) [Text] / I. S. Blekanov, N. A. Tarasov ; Rospatent. — № 2020662702 ; опубл. 16.10.2020 (Рос. Федерация).
4. *Blinova, O.* A hybrid model of complexity estimation: Evidence from Russian legal texts [Text] / O. Blinova, N. Tarasov // Frontiers in Artificial Intelligence. — 2022. — Vol. 5. — P. 1008530.
5. *Blekanov, I. S.* Transformer-based abstractive summarization for Reddit and Twitter: single posts vs. comment pools in three languages [Text] / I. S. Blekanov, N. Tarasov, S. S. Bodrunova // Future Internet. — 2022. — Vol. 14, № 3. — P. 69.
6. Topic detection based on sentence embeddings and agglomerative clustering with Markov moment [Text] / S. S. Bodrunova [et al.] // Future Internet. — 2020. — Vol. 12, № 9. — P. 144.
7. *Blinova, O. V.* Language Complexity across Sub-Styles and Genres in Legal Russian [Text] / O. V. Blinova, N. A. Tarasov // Research Result. Theoretical and Applied Linguistics. — 2023. — Vol. 9, № 2. — P. 73—96.
8. Data Encoding for Social Media: Comparing Twitter, Reddit, and Telegram [Text] / I. S. Blekanov [et al.] // Fifth Networks in the Global World Conference. — Springer. 2022. — P. 114—122.

9. Mapping opinion cumulation: topic modeling-based dynamic summarization of user discussions on social networks [Text] / I. S. Blekanov [et al.] // International Conference on Human-Computer Interaction. — Springer. 2023. — P. 25—40.

10. *Blinova, O. V.* Complexity of russian legal texts: assessment methods and language data [Text] / O. V. Blinova, N. A. Tarasov // Proceedings of the international conference "Corpus Linguistics-2021". — 2021. — P. 175.

11. Modeling lemma frequency bands for lexical complexity assessment of russian texts [Text] / O. Blinova [et al.] // Comput. Linguist. Intell. Technol. — 2020. — Vol. 19. — P. 76—92.

12. *Blinova, O. V.* Metrics of complexity of Russian legal texts: selection, use, initial assessment of effectiveness [Text] / O. V. Blinova, N. A. Tarasov // Computer linguistics and intellectual technologies: Based on the materials of the annual international conference “Dialogue”. Vol. 21, additional volume. — Russian Federation : Russian State Humanitarian University, 2022. — P. 1017—1028. — (Computational linguistics and intellegnt technologies).

13. *Reynolds, R. J.* Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories [Text] / R. J. Reynolds // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. — 2016. — P. 289—300.

14. *Collins-Thompson, K.* Computational assessment of text readability: a survey of current and future research [Text] / K. Collins-Thompson. — 2014.

15. *Crossley, S. A.* Moving beyond classic readability formulas: new methods and new models [Text] / S. A. Crossley, S. Skalicky, M. Dascalu // Journal of Research in Reading. — 2019. — Vol. 42, № 3/4. — P. 541—561.

16. *Benjamin, R. G.* Reconstructing readability: recent developments and recommendations in the analysis of text difficulty [Text] / R. G. Benjamin // , Educational Psychology Review. — 2012. — Vol. 24(1). — P. 63—88.

17. *Schwarm, S. E.* Reading level assessment using support vector machines and statistical language models [Text] / S. E. Schwarm, M. Ostendorf // 05) / ed. by P. of the 43rd Annual Meeting on Association for Computational Linguistics (acl. — 2005. — P. 523—530.

18. *Leroy, G.* The effect of word familiarity on actual and perceived text difficulty [Text] / G. Leroy, D. Kauchak // Journal of the American Medical Informatics Association. — 2014. — Vol. 21, e1. — e169—e172.

19. *Laposhina, A. N.* 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] [Text] / A. N. Laposhina. — 2017. — URL: http://www.dialog-21.ru/media/3993/ ; Proceedings of the International, Proceedings of the International Conference "Dialogue 2017" [Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Bekasovo.

20. *Ivanov, V. V.* Efficiency of text readability features in Russian academic texts [Text] / V. V. Ivanov, M. I. Solnyshkina, V. D. Solovyev // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2018. — Vol. 17. — P. 277—287.

21. *Sharoff, S.* Seeking needles in the web haystack: Finding texts suitable for language learners [Text] / S. Sharoff, S. Kurella, A. Hartley // Proceedings of 8th Teaching and Language Corpora Conference (TaLC-8. — 2008.

22. *Solovyev, V.* Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics [Text] / V. Solovyev, V. Ivanov, M. Solnyshkina // Journal of Intelligent & Fuzzy Systems. — 2018. — Vol. 34. — P. 3049—3058.

23. *Collins-Thompson, K.* Predicting Reading Difficulty with Statistical Language Models [Text] / K. Collins-Thompson, J. Callan // Journal of the American Society for Information Science and Technology. — 2005. — Vol. 56, № 13. — P. 1448—1462.

24. *Chen, X.* Characterizing Text Difficulty with Word Frequencies [Text] / X. Chen, W. D. Meurers // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. — 2016. — P. 84—94.

25. *Atkins, S.* Corpus Design Criteria [Text] / S. Atkins, J. Clear, N. Ostler // Literary and Linguistic Computing. — 1992. — Vol. 7. — P. 1—16.

26. *Biber, D.* Representativeness in Corpus Design [Text] / D. Biber // Literary and Linguistic Computing. — 1993. — Vol. 8, № 4. — P. 243—257.

27. *Brysbaert, M.* The Word Frequency Effect in Word Processing: An Updated Review, Current Directions in [Text] / M. Brysbaert, P. Mandera, E. Keuleers // Psychological Science. — 2018. — Vol. 27. — P. 45—50.

28. *Lyashevskaya, O. N.* Corpus Instruments for Russian Grammar Studies [Korpusnye instrumenty v grammaticheskih issledovanijah russkogo jazyka], Jazyki slavjanskoj kul'tury [Text] / O. N. Lyashevskaya. — 2016. — (Moscow).

29. *Schmitt, N.* Researching vocabulary: a vocabulary research manual [Text] / N. Schmitt. — Basingstoke, UK : Palgrave Macmillan, 2010.

30. *Zhao, Y.* The effect of lexical frequency and Lombard reflex on tone hyperarticulation [Text] / Y. Zhao, D. Jurafsky // Journal of Phonetics. — 2009. — Vol. 37. — P. 231—247.

31. Predictability effects on durations of content and function words in conversational English [Text] / A. Bell [et al.] // Journal of Memory and Language. — 2009. — Vol. 60. — P. 92—111.

32. ruTenTen11 [Text]. — URL: https://www.sketchengine.eu/rutenten-russian-corpus/.

33. The Sketch Engine: Ten Years On [Text] / A. Kilgarriff [et al.] // Lexicography. — 2014. — Vol. 1, Iss. 1. — P. 7—36.

34. *Russicum III, A. M.* / A. M. Russicum III. — URL: http://ucts.uniba.sk/aranea%5C_about/%5C_russicum.html.

35. *Benko, V.* Aranea: Yet Another Family of (Comparable) Web Corpora [Text] / V. Benko // Text, Speech and Dialogue. 17th International Conference, TSD 2014. Proceedings. LNCS 8655. Switzerland / ed. by P. Sojka [et al.]. — Springer International Publishing, 2014. — P. 257—264.

36. *Corpus, T.* An open-source corpus for machine learning [Text] / T. Corpus. — URL: https://tatianashavrina.github.io/taiga%5C_site/.

37. *Corpus, R. N.* / R. N. Corpus. — URL: http://www.ruscorpora.ru/new/.

38. *Lyashevskaya, O. N.* The frequency dictionary of modern Russian language [Častotnyj slovar' sovremennogo russkogo jazyka] [Text] / O. N. Lyashevskaya, S. A. Sharoff. — 2009. — URL: http://dict.ruslang.ru/freq.php ; csv-version.

39. *Kilgarriff, A.* Measures for corpus similarity and homogeneity [Text] / A. Kilgarriff, T. Rose // Proceedings of the Third Conference on Empirical Methods for Natural Language Processing. — Spain : Granada, 1998. — P. 46—52.

40. *Piperski, A. C.* Corpus Size and the Robustness of Measures of Corpus Distance, Computational Linguistics and Intellectual Technologies [Text] / A. C. Piperski // Dialogue 2018 / ed. by P. of the International Conference. — 2018. — P. 578—589.

41. *Gomaa, W. H.* A Survey of Text Similarity Approaches [Text] / W. H. Gomaa, A. A. Fahmy // International Journal of Computer Applications. — 2013. — Vol. 68. — P. 13—18.

42. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora, [Text] / M. Baroni [et al.] // Language Resources and Evaluation. — 2009. — Vol. 43. — P. 209—226.

43. Corpus Linguistics: An International Handbook [Text]. Vol. 2 / ed. by A. Lüdeling, M. Kytö. — Berlin, Boston : De Gruyter Mouton, 2009.

44. *Shaikevich, A. Y.* Measures of lexical similarity between frequency dictionaries [Mery leksicheskogo shodstva chastotnyh slovarej] [Text] / A. Y. Shaikevich // Corpus linguistics-2015 / ed. by P. of the Internationational Conference. — Saint Petersburg : Trudy mezhdunarodnoy nauchnoy konferentsii "Korpusnaya linguistica-2015"], 2015. — P. 434—442.

45. *Piperski, A.* Sum of Minimum Frequencies as a Measure of Corpus Similarity [Text] / A. Piperski // Presented at the Corpus Linguistics 2017, Birmingham. — 2017.

46. Subtlex-UK: A new and improved word frequency database for British English [Text] / W. J. B. Van Heuven [et al.] // Quarterly Journal of Experimental Psychology. — 2014. — Vol. 67. — P. 1176—1190.

47. *Sharoff, S.* Frequency Dictionary: Russian, Quasthoff U., Fiedler S., Hallsteindóttir E. (eds.), Frequency Dictinaries 9, Leipziger Universitätsverlag [Text] / S. Sharoff, D. Goldhahn, U. Quasthoff. — 2017.

48. *Jamieson, S.* Likert scales: how to abuse them [Text] / S. Jamieson // Medical Education. — 2004. — Vol. 38, № 12. — P. 1217—1218.

49. *Peter, M. T.* Legal Language [Text] / M. T. Peter. — Chicago, London : The University of Chicago Press, 1999.

50. *Heikki, E. S. M.* Comparative legal linguistics: language of law, Latin and modern lingua francas [Text] / E. S. M. Heikki // Ashgate Publishing, Ltd., Farnham, Surrey. — 2013. — Vol. 2 edition.

51. *Sol, A.-A.* On drafting, interpreting, and translating legal texts across languages and cultures [Text] / A.-A. Sol, Y. Ning // International Journal of Legal Discourse. — 2017. — Vol. 2, № 1. — P. 1—12.

52. *Vijay, K. B.* Cognitive structuring in legislative provisions [Text] / K. B. Vijay, G. John // Language and the Law. — 1994. — P. 136—155.

53. *linguistics, T. complexity: stages of study in domestic applied.* Solnyshkina, M. I. and Kiselnikov, A. S. [Text] / T. complexity: stages of study in domestic applied linguistics // Bulletin of Tomsk State University. — 2015. — Vol. 6, № 38. — P. 86—99.

54. *Juhan, T.* The development of statistical stylistics (a survey) [Text] / T. Juhan // Journal of Quantitative Linguistics. — 2017. — Vol. 11, № 1/2. — P. 141—151.

55. *Golub, I. B.* Stylistics of the Russian language [Text] / I. B. Golub. — Moscow : Rolf, 2001.

56. *Kozhina, M. N.* Russian language stylistics [Text] / M. N. Kozhina, L. Duskaeva, V. Salimovsky. — Moscow : Flint, Science, 2011.

57. *Druzhkin, K. Y.* Russian language readability metrics [Text] / K. Y. Druzhkin. — Higher School of Economics, Moscow : Master's final dissertation, 2016.

58. *Richard, C. W.* Plain English for lawyers [Text] / C. W. Richard, E. S. Amy. — 6th ed. — LLC, Durham, North Carolina : Carolina Academic Press, 2019.

59. *Robert, P. C.* Making legal language understandable: A psycholinguistic study of jury instructions [Text] / P. C. Robert, R. C. Veda // Columbia Law Review. — 1979. — Vol. 79, № 7. — P. 1306—1374.

60. *Marina, S.* Readability formula for russian texts: A modified version [Text] / S. Marina, I. Vladimir, S. Valery // Proceedings of the 17th Mexican International Conference on Artificial Intelligence. — MICAI 2018, 2018. — P. 132—145.

61. *Begtin, I.* Plainrussian.ru [Text] / I. Begtin. — 2016. — URL: https://github.com/ivbeg/readability.io.

62. *Milan, S.* Universal dependencies 2.5 models for UDPipe (2019-12-06) [Text] / S. Milan, S. Jana. — Faculty of Mathematics, Physics, Charles University : LINDAT/CLARIAH-CZ digital library at the Institute of Formal, Applied Linguistics (ÚFAL), 2019.

63. *Mikhail, K.* Morphological analyzer and generator for russian and ukrainian languages. // Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry Ignatov, and Valeri G [Text] / K. Mikhail // of Images, Social Networks and Texts, P / ed. by A. Labunets. — Cham : Springer International Publishing, 2015. — P. 320—332.

64. *Zhuravlev, A. F.* Experience in quantitative-typological research into varieties of oral speech [Text] / A. F. Zhuravlev // Varieties of urban oral speech. — 1988. — P. 84—150.

65. *Xiao, p. T.* Automatic genre classification via n-gr of part-of-speech tags [Text] / p. T. Xiao, C. Jing // AMS Procedia - Social and Behavioral Sciences, 198. — 2015. — P. 474—478.

66. *Antonova, A. Y.* Determination of style and genre characteristics of text collections based on part of speach compatibility. [Text] / A. Y. Antonova, E. S. Klyshinsky, E. V. Yagunova // Proceedings of the international conference "Corpus Linguistics-2011". — Saint Petersburg. SPbU, 2011. — P. 80—85.

67. *Nagel, O. V.* Word-formation mechanisms in the processes of perception, identification and use of language [Text] / O. V. Nagel. — Tomsk State University, Tomsk : diss. doc. Philol. sciences, 2017.

68. *Tianqi, C.* Xgboost: A scalable tree boosting system [Text] / C. Tianqi, G. Carlos // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — P, 2016. — P. 785—794.

69. Universal sentence encoder [Text] / C. Daniel [et al.]. — 2018. — arXiv preprint.

70. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms [Text] / J. Bergstra, D. Yamins, D. D. Cox, [et al.] // Proceedings of the 12th Python in science conference. Vol. 13. — Citeseer. 2013. — P. 20.

71. *Payam, R.* Lei Tang, and Huan Liu [Text] / R. Payam // Cross-validation. Encyclopedia of database systems. — 2016. — P. 1—7.

72. *Valery, S.* Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics [Text] / S. Valery, I. Vladimir, S. Marina // Journal of Intelligent Fuzzy Systems. — 2018. — Vol. 34. — P. 3049—3058.

73. Prediction of reading difficulty in Russian academic texts [Text] / S. Valery [et al.] // Journal of Intelligent Fuzzy Systems. — 2019. — Vol. 36. — P. 4553—4563.

74. *Deutsch, T.* Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories [Text] / T. Deutsch, M. Jasbi, S. Shieber // Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications / ed. by J. Tetreault [et al.]. — Association for Computational Linguistics, 2020. — P. 1—17.

75. *Szmrecsanyi, B.* Introduction: Linguistic complexity: Second Language Acquisition, indigenization, contact [Text] / B. Szmrecsanyi, B. Kortmann // Linguistic Complexity: Second Language Acquisition, Indigenization, Contact / ed. by B. Kortmann, B. Szmrecsanyi. — Berlin, Boston : De Gruyter, 2012. — P. 6—34.

76. *Dahl, ö.* The growth and maintenance of linguistic complexity [Text] / ö. Dahl. — Amsterdam : John Benjamins Publishing, 1993.

77. *Nichols, J.* Linguistic complexity : a comprehensive definition and survey [Text] / J. Nichols // Language complexity as an evolving variable / ed. by G. Sampson, D. Gil, P. Trudgill. — Oxford : Oxford University Press, 2009. — P. 110—125.

78. *Trudgill, P.* Sociolinguistic typology: Social determinants of linguistic complexity [Text] / P. Trudgill. — Oxford : Oxford University Press, 2011.

79. *McWhorter, J.* The worlds simplest grammars are creole grammar [Text] / J. McWhorter // Linguistic Typology. — 2001. — Vol. 5, № 2/3. — P. 125—166. — URL: https://www.degruyter.com/document/doi/10.1515/lity.2001.001/html.

80. *Frazier, L.* Syntactic complexity [Text] / L. Frazier // Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives / ed. by D. R. Dowty, L. Karttunen, A. M. Zwicky. — Cambridge : Cambridge University Press, 1985. — P. 129—189.

81. *Collins-Thompson, K.* Computational assessment of text readability: a survey of current and future research [Text] / K. Collins-Thompson // Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics. — 2014. — Vol. 165, № 2. — P. 97—135. — URL: https://benjamins.com/catalog/itl.165.2.01col.

82. Text complexity as interdisciplinary problem [Text] / M. Solnyshkina [et al.] // Voprosy Kognitivnoy Lingvistiki. — 2022. — № 1. — P. 18—39. — URL: http://vcl.ralk.info/issues/2022/vypusk-1-2022/slozhnost-teksta-kak-mezhdistsiplinarnaya-problema.html.

83. *Tiersma, P. M.* Legal Language [Text] / P. M. Tiersma. — Chicago, London : The University of Chicago Press, 1999.

84. *Azuelos-Atias, S.* On drafting, interpreting, and translating legal texts across languages and cultures [Text] / S. Azuelos-Atias, N. Ye // International Journal of Legal Discourse. — 2017. — Vol. 2, № 1. — P. 1—12. — URL: https://www.degruyter.com/document/doi/10.1515/ijld-2017-1000/html.

85. *Wydick, R. C.* Plain English for lawyers [Text] / R. C. Wydick, A. E. Sloan. — Durham, North Carolina : Carolina Academic Press, LLC, 2019.

86. *Dmitrieva, A. V.* "The art of legal writing": A quantitative analysis of Russian Constitutional Court rulings [Text] / A. V. Dmitrieva // Sravnitel'noe konstitutsionnoe obozrenie. — 2017. — Vol. 118, № 3. — P. 125—133. — URL: https://sko-journal.ru/catalog/sko-3-118-2017/iskusstvo-yuridicheskogo-pisma-kolichestvennyj-analiz-reshenij-konstitutsionnogo-suda-rossii/.

87. *Oborneva, I. V.* Automation of text perception quality assessments [Text] / I. V. Oborneva // Vestnik Moskovskogo gorodskogo pedagogicheskogo universiteta. — 2005. — № 5. — P. 86—91. — URL: https://www.elibrary.ru/item.asp?id=12804809.

88. *Kuchakov, R.* The complexity of legal acts in Russia: Lexical and syntactic quality of texts: analytic note [Text] / R. Kuchakov, D. Savel'ev. — Saint Petersburg : European University at Saint Petersburg, 2018.

89. *Savel'ev, D.* Decisions of arbitration courts of Russian Federation: lexical and syntactic quality of texts, analytic note [Text] / D. Savel'ev, R. Kuchakov. — Saint Petersburg : European University at Saint Petersburg, 2019.

90. Complexity of Russian Laws. The Experience of Syntactic Analysis [Text] / A. Knutov [et al.]. — Moscow : HSE University Publishing House, 2020.

91. *Collins-Thompson, K.* A language modeling approach to predicting reading difficulty [Text] / K. Collins-Thompson, J. P. Callan // Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004. — 2004. — P. 193—200.

92. A comparison of features for automatic readability assessment [Text] / L. Feng [et al.] // COLING'10: Proceedings of the 23rd International Conference on Computational Linguistics / ed. by C. 2. O. Committee. — International Committee on Computational Linguistics, 2010. — P. 276—284.

93. *Xia, M.* Text readability assessment for second language learners [Text] / M. Xia, E. Kochmar, T. Briscoe // arXiv preprint arXiv:1906.07580. — 2019.

94. Automated Assessment of Language Proficiency on German Data [Text] / E. Szügyi [et al.] // KONVENS. — 2019. — P. 41—50.

95. Automatic classification of text complexity [Text] / V. Santucci [et al.] // Applied Sciences. — 2020. — Vol. 10, № 20. — P. 7285.

96. *Lyashevskaya, O.* Automated assessment of learner text complexity [Text] / O. Lyashevskaya, I. Panteleeva, O. Vinogradova // Assessing Writing. — 2021. — Vol. 49. — P. 100529.

97. *Staudemeyer, R. C.* Understanding LSTM–a tutorial into long short-term memory recurrent neural networks [Text] / R. C. Staudemeyer, E. R. Morris // arXiv preprint arXiv:1909.09586. — 2019.

98. *Morozov, D. A.* Text complexity and linguistic features: Their correlation in English and Russian [Text] / D. A. Morozov, A. V. Glazkova, B. L. Iomdin // Russian Journal of Linguistics. — 2022. — Vol. 26, № 2. — P. 426—448.

99. *Sharoff, S. A.* What neural networks know about linguistic complexity [Text] / S. A. Sharoff // Russian Journal of Linguistics. — 2022. — Vol. 26, № 2. — P. 371—390.

100. Efficient estimation of word representations in vector space [Text] / T. Mikolov [et al.] // arXiv preprint arXiv:1301.3781. — 2013.

101. *Pennington, J.* Glove: Global vectors for word representation [Text] / J. Pennington, R. Socher, C. D. Manning // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532—1543.

102. Enriching word vectors with subword information [Text] / P. Bojanowski [et al.] // Transactions of the association for computational linguistics. — 2017. — Vol. 5. — P. 135—146.

103. *Bosco, G. L.* A neural network model for the evaluation of text complexity in Italian language: a representation point of view [Text] / G. L. Bosco, G. Pilato, D. Schicchi // Procedia computer science. — 2018. — Vol. 145. — P. 464—470.

104. A Transfer Learning Based Model for Text Readability Assessment in German [Text] / S. Mohtaj [et al.] // arXiv preprint arXiv:2207.06265. — 2022.

105. Bert: Pre-training of deep bidirectional transformers for language understanding [Text] / J. Devlin [et al.] // arXiv preprint arXiv:1810.04805. — 2018.

106. *Dmitrieva, A.* A Comparative Study of Educational Texts for Native, Foreign, and Bilingual Young Speakers of Russian: Are Simplified Texts Equally Simple? [Text] / A. Dmitrieva, A. Laposhina, M. Lebedeva // Frontiers in Psychology. — 2021. — Vol. 12. — URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.703690.

107. Consultant Plus: Legal Reference System [Text]. — 2022. — URL: http://www.consultant.ru ; Accessed August 30, 2022.

108. Garant: Legal information portal [Text]. — 2022. — URL: https://www.garant.ru/ ; Accessed August 30, 2022.

109. *Ivanov, V.* Efficiency of text readability features in Russian academic texts [Text] / V. Ivanov, M. Solnyshkina, V. Solovyev // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii 2018 (Computational Linguistics and Intellectual Technologies 2018). — 2018. — Vol. 17, № 24. — P. 284—293. — URL: https://www.dialog-21.ru/media/4302/ivanovvv.pdf.

110. Using Universal Dependencies in cross-linguistic complexity research [Text] / A. Berdicevskis [et al.] // Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). — Association for Computational Linguistics, 2018. — P. 8—17.

111. *Korobov, M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages [Text] / M. Korobov // Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science / ed. by M. Y. Khachay [et al.]. — Springer International Publishing, 2015. — P. 320—332.

112. CoNLL 2018 Shared Task [Text]. — 2018. — URL: https : / / universaldependencies . org / conll18 / evaluation . html ; Accessed August 30, 2022.

113. *Druzhkin, K.* Readability metrics for Russian: master's thesis [Text] / K. Druzhkin. — Moscow : Higher School of Economics, 2016.

114. *Benjamin, R.* Reconstructing readability: recent developments and recommendations in the analysis of text difficulty [Text] / R. Benjamin // Educational Psychology Review. — 2012. — № 24. — P. 63—88. — URL: https://link.springer.com/article/10.1007/s10648-011-9181-8.

115. *Solnyshkina, M.* Readability Formula for Russian Texts: A Modified Version [Text] / M. Solnyshkina, V. Ivanov, V. Solovyev // Advances in Computational Intelligence. MICAI 2018. Lecture Notes in Computer Science. — 2018. — Vol. 11289. — P. 132—145. — URL: https://link.springer.com/chapter/10.1007/978-3-030-04497-8%5C_11.

116. *Begtin, I.* PlainRussian [Text] / I. Begtin. — 2016. — URL: https://github.com/ivbeg/readability.io.

117. *Straka, M.* Universal Dependencies 2.5 Models for UDPipe [Text] / M. Straka, J. Straková. — 2016. — URL: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131.

118. *Zhuravlev, A. F.* Experience of quantitative-typological study of varieties of oral speech [Text] / A. F. Zhuravlev // Raznovidnosti gorodskoi ustnoi rechi. Sbornik nauchnykh trudov / ed. by D. Shmelev, E. Zemskaja. — Moscow : Nauka, 1988. — P. 84—150.

119. Formation of a model of compatibility of Russian words and the study of its propertie [Text] / J. S. Klyshinskij [et al.]. — Moscow : Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, 2013.

120. *Antonova, A. J.* Determination of stylistic and genre characteristics of text collections based on part-of-speech compatibility [Text] / A. J. Antonova, E. S. Klyshinsky, E. V. Jagunova // Trudy mezhdunarodnoj konferencii "Korpusnaja lingvistika–2011" / ed. by V. P. Zaharov. — Saint Petersburg State University, 2011. — P. 80—85.

121. *Dobrego, A.* Processing of static and dynamic texts: an eye-tracking study of Russian [Text] / A. Dobrego, T. Petrova // 3rd International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2016. Vol. 1.1 / ed. by S. editorial board. — STEF92 Technology, 2016. — P. 991—998.

122. *Nagel', O. V.* Word-formation mechanisms in the processes of perception, identification, and use of language: author's abstract of the doctor's thesis [Text] / O. V. Nagel'. — Tomsk : National Research Tomsk State University, 2017.

123. *Kyle, K.* Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices [Text] / K. Kyle, S. A. Crossley // The Modern Language Journal. — 2018. — Vol. 102, № 2. — P. 333—349. — URL: https://onlinelibrary.wiley.com/doi/10.1111/modl.12468.

124. *Biber, D.* Grammatical Complexity in Academic English. Linguistic Change in Writing [Text] / D. Biber, B. Gray. — Cambridge : Cambridge University Press, 2016.

125. *Ljashevskaja, O. N.* On Determining the Complexity of Russian Texts [Text] / O. N. Ljashevskaja // XVII Aprel'skaia mezhdunarodnaia nauchnaia konferentsiia po problemam razvitiia ekonomiki i obshchestva: v 4 kn. / ed. by E. G. Jasin. — HSE University Publishing House, 1996. — P. 408—419.

126. *Bentz, C.* Zipf's law of abbreviation as a language universal [Text] / C. Bentz, R. Ferrer-i-Cancho // Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics / ed. by C. Bentz, G. Jäger, I. Yanovich. — University of Tübingen, 2016. — P. 1—4.

127. What happens to bert embeddings during fine-tuning? [Text] / A. Merchant [et al.] // arXiv preprint arXiv:2004.14448. — 2020.

128. *Kuratov, Y.* Adaptation of deep bidirectional multilingual transformers for russian language [Text] / Y. Kuratov, M. Arkhipov // arXiv preprint arXiv:1905.07213. — 2019.

129. Huggingface's transformers: State-of-the-art natural language processing [Text] / T. Wolf [et al.] // arXiv preprint arXiv:1910.03771. — 2019.

130. *Loshchilov, I.* Decoupled weight decay regularization [Text] / I. Loshchilov, F. Hutter // arXiv preprint arXiv:1711.05101. — 2017.

131. *Chen, T.* Xgboost: A scalable tree boosting system [Text] / T. Chen, C. Guestrin // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — 2016. — P. 785—794.

132. *Blinova, O.* Decisions of Russian Constitutional Court: Lexical Complexity Analysis in Shallow Diachrony [Text] / O. Blinova, S. Belov, M. Revazov // CEUR Workshop Proceedings. Vol-2813. Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, Russia 17-20 June 2020 / ed. by R. Bolgov, A. V. Chugunov, A. E. Voiskounsky. — The name of the publisher, 2020. — P. 61—74.

133. *Assy, R.* Can the Law Speak Directly to its Subjects? The Limitation of Plain Language [Text] / R. Assy // Journal of Law and Society. — 2013. — Vol. 38, № 3. — P. 376—404. — URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6478.2011.00549.x.

134. *S., G.-R.* Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study // Łódź Studies in Language 22 [Text] / G.-R. S. // Berlin, Peter Lang Verlag: — 2012. — P. 280.

135. *Orts, M. Á.* Power and Complexity in Legal Genres: Unveiling Insurance Policies and Arbitration Rules [Text] / M. Á. Orts // International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique. — 2015. — Vol. 28. — P. 485—505.

136. *Martínez, E.* Poor writing, not specialized concepts, drives processing difficulty in legal language // Cognition [Text] / E. Martínez, F. Mollica, E. Gibson. — 224, 2022. — (Vol).

137. *Venturi, G.* Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach [Text] / G. Venturi // IALF Porto. — 2012. — P. 138—156.

138. *McKinley, J.* Text analysis [Text] / J. McKinley, R. (Heath // The Routledge Handbook of Research Methods in Applied Linguistics. — 2019. — P. 453—463.

139. *Swales, J. M.* English in Academic and Research Settings [Text] / J. M. Swales. — Cambridge, Cambridge University Press, 1990.

140. *Bhatia, V. K.* Genre: Language use in Professional Settings. Applied linguistics and language study [Text] / V. K. Bhatia. — London : Routledge, Taylor & Francis, 2013.

141. *Durant, A.* Legal Genres // Language and Law: A Resource Book for Students [Text] / A. Durant, J. H. Leung // : Routledge / ed. by R. E. L. Introductions. — London : Taylor & Francis, 2016. — P. 11—15.

142. *Tessuto, G.* Investigating English Legal Genres in Academic and Professional Contexts [Text] / G. Tessuto // Cambridge: Cambridge Scholars Publishing. — 2012. — Vol. 315 p.

143. *Bhatia, V. K.* An applied discourse analysis of English legislative writing [Text] / V. K. Bhatia // Birmingham: University of Aston in Birmingham. — 1983. — Vol. 145 p.

144. *Kurzon, D.* How Lawyers Tell their Tales: Narrative Aspects of a Lawyer's Brief [Text] / D. Kurzon // Poetics. — 1985. — Vol. 14. — P. 467—481.

145. *Tiersma, P. M.* The Language of Offer and Acceptance: Speech Acts and the Question of Intent [Text] / P. M. Tiersma // California Law Review. — 1986. — Vol. 74. — P. 189—232.

146. *Trosborg, A.* An analysis of legal speech acts in English Contract Law. "It is hereby performed." // HERMES - Journal of Language and Communication in Business [Text] / A. Trosborg // Vol. — 1991. — Vol. 4. — P. 65—90.

147. *Trosborg, A.* Statutes and contracts: An analysis of legal speech acts in the English language of the law [Text] / A. Trosborg // Journal of Pragmatics. — 1995. — Vol. 23. — P. 31—53.

148. *Howe, P. M.* The problem of the problem question in English for academic legal purposes // English for Specific Purposes [Text] / P. M. Howe // №. — 1990. — Vol. 9. — P. 215—236.

149. *M., T. R. A.* Subject Specific Literacy and Genre Theory // Australian Review of Applied Linguistics [Text] / T. R. A. M. // Legal English. — 1993. — Vol. 16. — P. 86—122.

150. *Savelyev, D. A.* Study of the complexity of sentences that make up the texts of legal acts of the authorities of the Russian Federation [Text] / D. A. Savelyev // Right. Journal of the Higher School of Economics. — 2020. — Vol. Vol. 1. C. — P. 50—74.

151. *Goźdź-Roszkowski, S.* Legal terms in context: phraseological variation across genres // Evidence-Based LSP: Translation, Text and Terminology, Linguistic Insights: Studies in Language and Communication [Text] / S. Goźdź-Roszkowski // Bern: Peter Lang AG. — 2007. — P. 455—470.

152. *Dell'Orletta, F.* Genre-oriented Readability Assessment: a Case Study // Proceedings of the Workshop on Speech and Language Processing Tools in Education [Text] / F. Dell'Orletta, G. Venturi, S. Montemagni // The COLING. — 2012. — Vol. 2012 Organizing Committee, Mumbai. — P. 91—98.

153. Continent [Text]. — 2023. — URL: https://continent-online.com/.

154. Techexpert [Text]. — 2023. — URL: https://cntd.ru/about/network.

155. *Borisov, A. B.* / A. B. Borisov // Large legal dictionary. — 2010. — P. 848.

156. *Dodonov, V.* Another large legal dictionary [Text] / V. Dodonov // M.: Scientific Publishing Center INFRA-M. — 2001. — P. 780.

Figures

Tables