UDC 81'322 DOI: 10.18413/2313-8912-2023-9-2-0-5

Olga V. Blinova[1]
Nikita A. Tarasov[2]

Language complexity across sub-styles and genres in legal Russian

[1] Saint Petersburg University
7-9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
National Research University Higher School of Economics
16 Soyuza Pechatnikov St., Saint Petersburg, 190121, Russia
E-mail: o.blinova@spbu.ru

[2] Saint Petersburg University
7-9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
E-mail: nkt.tarasov@yandex.ru

Received 21 January 2023; accepted 25 April 2023; published 30 June 2023

Acknowledgements. The research is supported by the Russian Science Foundation, Project #19-18-00525 “Understanding official Russian: The legal and linguistic issues”.

Abstract. The purpose of the paper is to find out the differences in linguistic complexity between legal documents, opposed by domain, sub-style and genre. The authors explore the large and diverse corpus of Russian legal texts and compare (1) international documents and documents of national law, (2) documents of the three sub-styles (administrative, legislative and justiciary), and (3) texts of different genres within sub-styles. To obtain complexity scores, an automatic model is used whose modules are capable of predicting complexity either by using the fine-tuned ruBERT model, or by using 133 language metrics, or in a hybrid way. The paper analyzes a dataset consisting of 43,804 documents, 118,768,028 words. National law documents are classified into three sub-styles. In addition, each document is characterized according to the genre and to the issuing body. Thus, 68 genres were identified. All documents were assigned complexity scores ranging from “0” to “12”. The vast majority of all documents were scored as maximally complex. The hybrid model assigned a complexity level of “12” to 97.1% of administrative sub-style documents, 94.5% of legislative sub-style documents, and 99.7% of judicial sub-style documents of national law. For all international law documents, the proportion of documents with a level of complexity of “12” is 94.1%. The set of legislative sub-style texts is the most varied in complexity. On average, the most complex documents in the dataset are of justiciary sub-style ones. Linguistic features successfully contrast international and national documents, as well as legislative and justiciary sub-styles. When comparing documents by genre, the authors interpreted only the average values of the 22 syntactic metrics. In general, a comparison of the genre-based document groups showed that it is not the genre itself that may be decisive for the complexity score, but the issuing body.

Keywords: Language Complexity; Legal Russian; Complexity Assessment Model; Sub-styles; Genre Analysis; Administrative sub-style documents; Legislative sub-style documents; Justiciary Sub-style Documents

How to cite: Blinova, O. V. and Tarasov, N. A. (2023). Language complexity across sub-styles and genres in legal Russian, *Research Result. Theoretical and Applied Linguistics*, 9 (2), 73-96. DOI: 10.18413/2313-8912-2023-9-2-0-5

УДК 81'322 **DOI: 10.18413/2313-8912-2023-9-2-0-5**

Блинова О. В.[1]
Тарасов Н. А.[2]

Языковая сложность русских юридических подстилей и жанров

[1] Санкт-Петербургский государственный университет
Университетская набережная, 7-9, Санкт-Петербург, 199034, Россия
Национальный исследовательский университет «Высшая школа экономики»
ул. Союза Печатников, 16, Санкт-Петербург, 190121, Россия
E-mail: o.blinova@spbu.ru

[2] Санкт-Петербургский государственный университет
Университетская набережная, 7-9, Санкт-Петербург, 199034, Россия
E-mail: nkt.tarasov@yandex.ru

*Статья поступила 21 января 2023 г.; принята 25 апреля 2023 г.;
опубликована 30 июня 2023 г.*

Информация об источниках финансирования или грантах: исследование поддержано Российским научным фондом, проект № 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика».

Аннотация. Цель статьи – выяснить, каковы отличия в языковой сложности между юридическими документами, противопоставленными по областям права, подстилям и жанрам. Авторы рассматривают обширный и разнообразный корпус русских юридических текстов и сравнивают (1) международные документы и документы национального права, (2) документы трёх подстилей (административного, законодательного и юрисдикционного), а также (3) тексты разных жанров внутри подстилей. Для получения оценок сложности используется автоматическая модель, модули которой способны предсказывать сложность либо с использованием дообученной языковой модели ruBERT, либо с использованием 133 языковых метрик, либо гибридным образом. Анализируется коллекция, состоящая из 43 804 документов и включающая 118 768 028 слов. Документы национального права классифицированы по трём подстилям, каждый документ охарактеризован в соответствии с жанром и издавшим его органом. Таким образом выделены 68 жанров. Всем документам присвоены оценки сложности в диапазоне от «0» до «12». Выяснено, что подавляющее большинство всех документов оценивается как максимально сложные. Так, гибридная модель присваивает класс сложности «12» 97,1% документов административного, 94,5% документов законодательного и 99,7% документов судебного подстиля. По отношению ко всем документам международного права доля документов с уровнем сложности «12» составляет 94,1%. Набор текстов законодательного подстиля является самым разнообразным по сложности. В среднем самые сложные документы в исследуемом наборе данных относятся к юрисдикционному подстилю. Лингвистические признаки успешно противопоставляют международные и национальные документы, а также документы законодательного и юрисдикционного подстилей. При сравнении документов по жанрам авторы интерпретировали средние значения 22 синтаксических метрик. В целом сравнение жанровых групп показывает, что решающее значение для оценки сложности может иметь не собственно жанр, а издавший документ государственный орган.

Ключевые слова: Языковая сложность; Русские правовые тексты; Модель оценки сложности; Подстили; Анализ жанров; Документы административного подстиля; Документы законодательного подстиля; Документы юрисдикционного подстиля

Информация для цитирования: Блинова О. В., Тарасов Н. А. Языковая сложность русских юридических подстилей и жанров // Научный результат. Вопросы теоретической и прикладной лингвистики. 2023. Т. 9. № 2. С. 73-96. DOI: 10.18413/2313-8912-2023-9-2-0-5

Introduction

This paper 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**” (Goźdź-Roszkowski, 2012: 11).

(Mattila, 2013) 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. (Orts, 2015) on two internationally used documents, (Martínez et al., 2022) on contracts, (Venturi, 2012) on different sub-varieties of legal language.

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

We use the approaches to classifying styles, sub-styles and genres, proposed by Russian **functional stylistics**. 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., (Kozhina et al., 2011: 329). In addition, **diplomatic** sub-style is distinguished. The documents of this sub-style regulate legal relations between states.[1]

Firstly, in this paper, we distinguish between **documents of national law** and **international legal documents**. 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, we only consider **synchronous documents**. The notion of synchronicity is formalized for the purposes of this paper as follows. “Synchronous” is

[1] There are also other classifications of official business sub-styles. Thus, B. S. Schwarzkopf (1996) speaks of three sub-styles (“bureaucratic-business”, “legal” and “diplomatic” ones), G. Ya. Solganik (2003) distinguishes two sub-styles (“official-documentary” and “casual-business” ones).

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, we are analyzing documents of the Russian Federation, 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, we look at particular **legal genres**. 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 paper, see section 2.1 below.

Fourthly, we consider only **written legal genres**; oral genres remain outside the scope of this paper.

This paper is structured as follows. The 1st section provides a brief literature review. The 2nd section includes three subsections: Subsection 2.1 describes how we collected the genre-diverse legal corpus; in Subsection 2.2 we characterize the dataset to be analyzed, and in Subsection 2.3 we briefly report our complexity assessment model. The 3rd section presents the results of the complexity analysis, a comparison of international documents and documents of national law, a comparison of the three sub-styles, and a comparison of documents of various genres for each of the sub-styles.

1. Literature review

1.1. 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 (Wang, 2019: 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" (Swales, 1990: 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" (Bhatia, 2013: 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, (Mattila, 2013) 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 (Durant and Leung, 2016: 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 (Tessuto, 2012: 13). There are research works on legislation and legal genres by Bhatia (1983, 2013), on lawyers' briefs by Kurzon (1985), on contracts by Tiersma (1986) and Trosborg (1991), on legislative texts and contracts by Trosborg (1995), on professional argumentation of lawyers by Howe (1990), on apprenticeship into academic discourse community and degrees of linguistic intricacy by Iedema (1993).

1.2. Complexity studies

There are plenty of research works related to the language complexity analysis, for an overview see e.g. (Solnyshkina et al., 2022). 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) on the complexity of Judgments of the RF Constitutional Court, and other research works, which we discuss below. In the paper (Dmitrieva, 2017), complexity was evaluated using a single readability formula. (Saveliev, Kuchakov, 2019) 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" (Kuchakov and Saveliev, 2018). 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" (Kuchakov and Saveliev, 2018).

The most genre-diverse sample of Russian legal texts has been analyzed in (Saveliev, 2020); 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".[2] 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 (Knutov et al., 2020), (Kuchakov, Saveliev, 2018), 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,

[2] Ukaz Prezidenta Rossijskoj Federacii "Ob obshhepravovom klassifikatore otraslej zakonodatel'stva" [Decree of the President of the Russian Federation "On branches of legislation"] (1993). Sobranie aktov Prezidenta i Pravitel'stva Rossijskoj Federacii ot 1993 g. [Collection of Acts of the President and Government of the Russian Federation of 1993], № 51, st. 4936. URL: https://docs.cntd.ru/document/901752675 (Accessed 15 January 2023). *(In Russian)*

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, (Goźdź-Roszkowski, 2007) on legal terminology. (Dell'Orletta et al., 2012) 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".

2. Materials and methods

2.1. Legal documents

In order to understand which documents are to be included in the legal corpus, we looked at the taxonomies from the Russian legal databases and documentation databases, namely (Consultant Plus), (Garant), (Continent), (Techexpert).[3] Based on this information, a preliminary list of document types was generated, containing 591 items (further – "list-591"). To evaluate this list, we turned to legal experts and conducted an experiment 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*?". We then assessed the consistency of responses 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, 2010) and (Dodonov et al., 2001). 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 our 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. We then used the combined list of legal "document types" (612 items) to obtain the texts of documents from legal database sites and sites of state authorities.

2.2. Analyzing data

Using the list of document types (see the previous section), we obtained legal documents and formed a text collection. Then we normalized the names of documents from this text collection and thus received a list of genres, consisting of 306 items. We divided all genres into the following categories: international documents vs. documents of national law (**administrative sub-style documents, legislative sub-style documents,**

[3] Consultant Plus: Legal Reference System. URL: http://www.consultant.ru (Accessed 15 January 2023). Garant: Legal information portal. URL: https://www.garant.ru/ (Accessed 15. January 2023). Information system "Continent". URL: https://continent-online.com/ (Accessed 15 January 2023). Information network "Techexpert". URL: https://cntd.ru/about/network (Accessed 15 January 2023).

[4] Obscherossiisky klassifikator upravlencheskoy dokumentatsii OK 011-93 [Russian Classification of Management Documentation OK 011-93], 1994. URL: https://docs.cntd.ru/document/9035738 (Accessed 15 January 2023). *(In Russian)*

and justiciary sub-style documents; **further we will refer to the corresponding documents using acronyms ASSDs, LSSDs and JSSDs**). In the next step we selected the genres to be analyzed in this paper (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 1. 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.

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.

Table 1. Genres of National Law Documents
Таблица 1. Жанры документов национального права

SS	#Genres	List of Genres	#Documents	#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

The International Law dataset consists of 1,617 texts, 6,400,239 words, includes international agreements, conventions, decrees, and judgments of international courts.

2.3. Complexity Estimation Model

Our complexity model is described in detail in (Blinova and Tarasov, 2022), so here we will limit ourselves to its brief specification. The model has been composed in two main stages.

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 our 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 to our 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. Collecting and labeling for highly complex examples of legal texts are the subjects of currently ongoing work.

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:

1. basic metrics, traditionally used in the tasks of readability assessment;
2. readability formulas, adapted for the Russian language;
3. words of various part-of-speech classes;
4. part-of-speech n-grams;
5. general-language frequency characteristics of text lemmas;
6. word-formation patterns;
7. separate grammes;
8. lexical and semantic features, multi-word expressions;
9. syntactic features
10. 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.

We have 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, we have 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 is available at https://github.com/PlainDocument and can be used as a hybrid model, feature based model or language modeling-based model.

3. Results and Discussion

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

Table 2, Table 3, and Table 4 below present the results of language complexity estimation for national law documents (ASSDs, LSSDs and JSSDs), and international law documents. Table 2 shows the results of the hybrid model, Table 3 shows the ruBERT predictions, and Table 4 shows the metrics-based complexity predictions.

Table 2. Hybrid Predictions
Таблица 2. Предсказания гибридной модели

Complexity Level	National Law Documents			
	Administrative Sub-style Documents	Legislative Sub-style Documents	Justiciary Sub-style Documents	International Law Documents
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

Table 3. RuBERT Predictions
Таблица 3. Предсказания RuBERT

Complexity Level	National Law Documents			
	Administrative Sub-style Documents	Legislative sub-style Documents	Justiciary Sub-style Documents	International Law Documents
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 4. Metrics Predictions
Таблица 4. Предсказания модели на метриках

Complexity Level	**National Law Documents**			
	Administrative Sub-style Documents	**Legislative Sub-style Documents**	**Justiciary Sub-style Documents**	**International Law Documents**
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

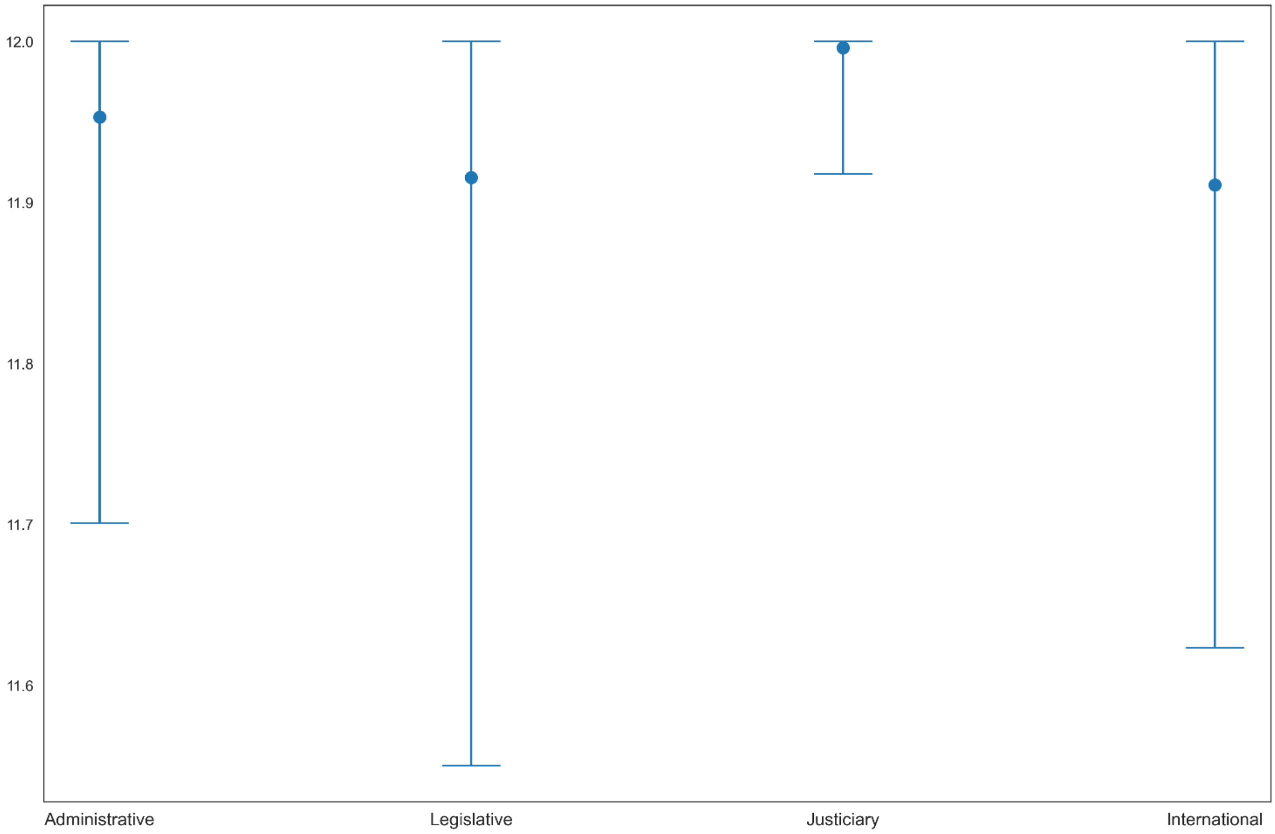
The results show that the vast majority of all documents in all of our large classes are rated by all models as maximally complex. For instance, if we take a closer look at the results of the hybrid model (see Table 2), **complexity class “12” includes 97.1% of administrative sub-style documents, 94.5% 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. Let us give an explanation of how the models work on a complexity level of “0”, which we actually did not expect to see in our 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 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. 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.

[5] See the full text on the Information network “Techexpert”.
URL: https://docs.cntd.ru/document/901810772

Figure 1. Mean Values of Complexity (Hybrid Predictions)
Рисунок 1. Средние значения сложности (Предсказания гибридной модели)

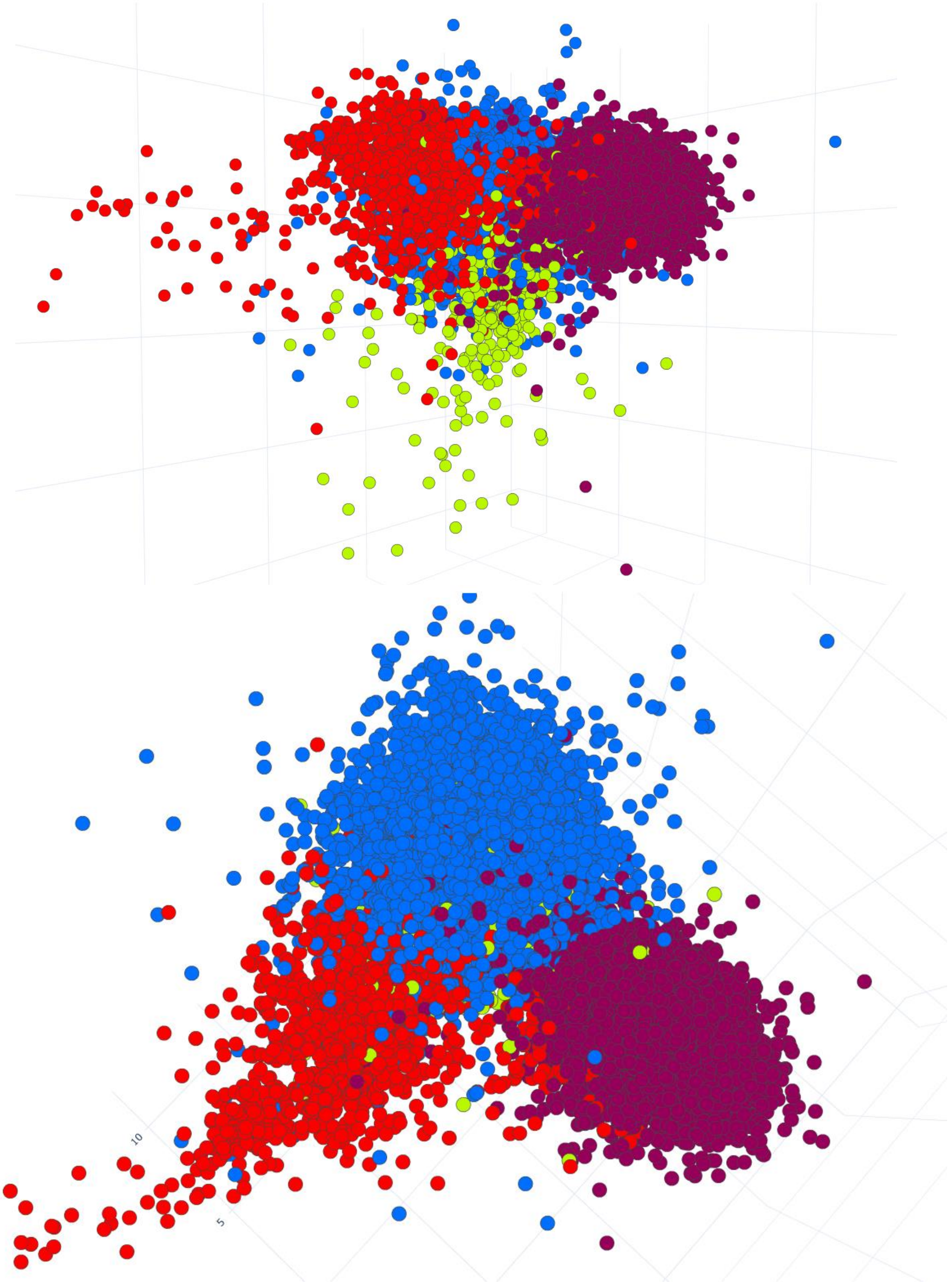


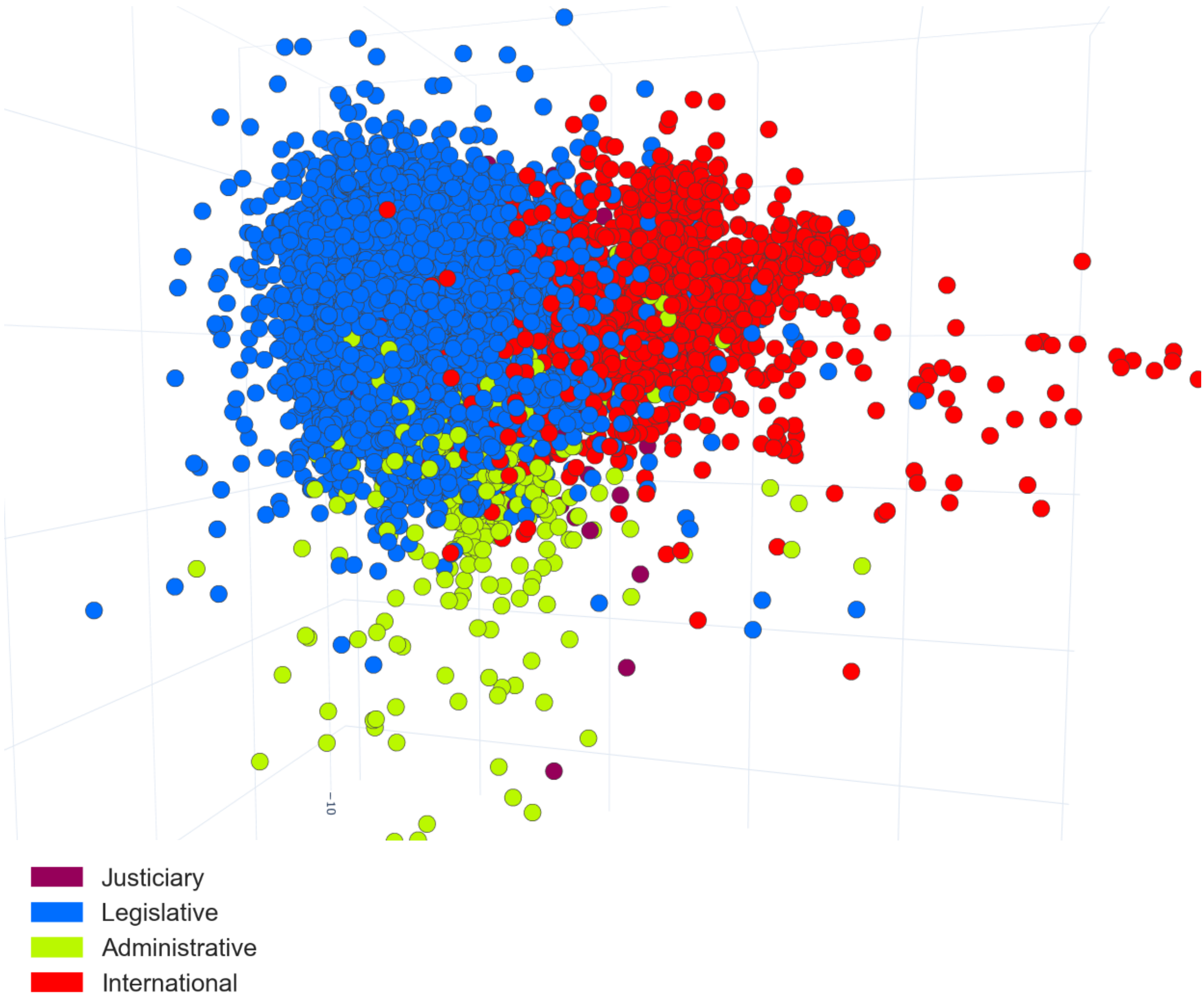
The visualization confirms that the most complex documents in the studied dataset are JSSDs.

Linear Discriminant Analysis (LDA) was performed to reduce the dimensions of the feature vectors from 133 language parameters down to 3. Fig. 2 shows the visualization of sub-styles and statuses using the reduced vectors for each document.

Fig. 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.

Figure 2. Documents Comparison using LDA for Dimensionality Reduction (three projections)
Рисунок 2. Сравнение документов с использованием LDA для уменьшения размерности (три проекции)



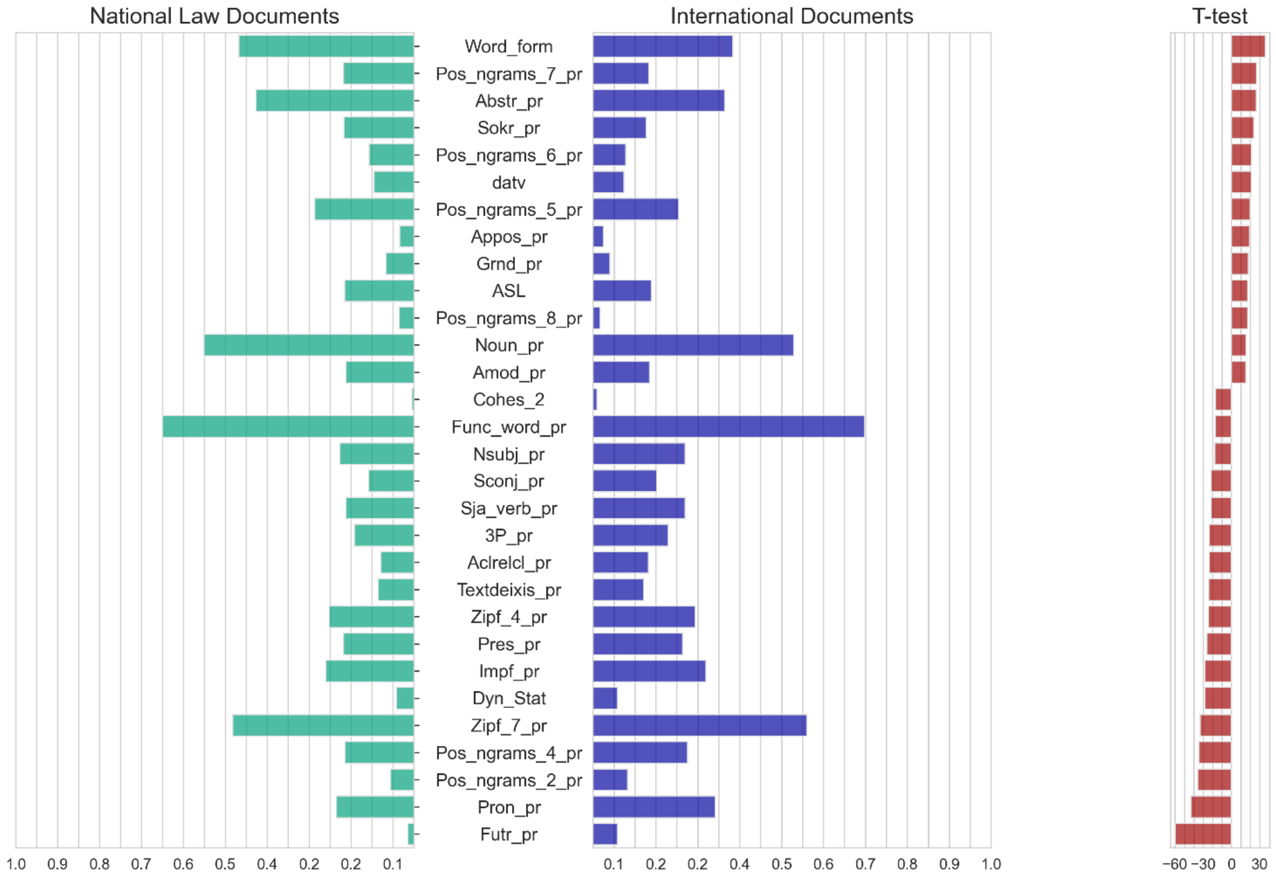


For a more detailed comparison of documents by the status, we analyzed the mean values of linguistic metrics. To compare these values between national law documents and international ones a t-test was performed. It has been found that for Bonferroni adjusted p-values less than 0.05, we can reject the null-hypothesis (equal mean values) 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. 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. A **Full list of linguistic metrics** is given on the RSF-funded project's website https://www.plaindocument.org/corpora.

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.

Figure 3. Mean Values of Linguistic Metrics in Documents by Status
Рисунок 3. Средние значения лингвистических метрик в документах по статусу



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”.

3.2. 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 (see Appendix to this paper). Averages were calculated after the min-max normalization of each feature. Fig. 4, 5 and 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 allow us to give a meaningful interpretation of the data obtained, since we did not get a very diverse distribution of domestic documents according to the complexity scores (see section 3.1 above for more information). Thus, **we make generalizations based on syntactic features**, because we consider them the most revealing in assessing text complexity.

Figure 4. Genres' Complexity within Administrative Sub-style
Рисунок 4. Сложность жанров административного подстиля

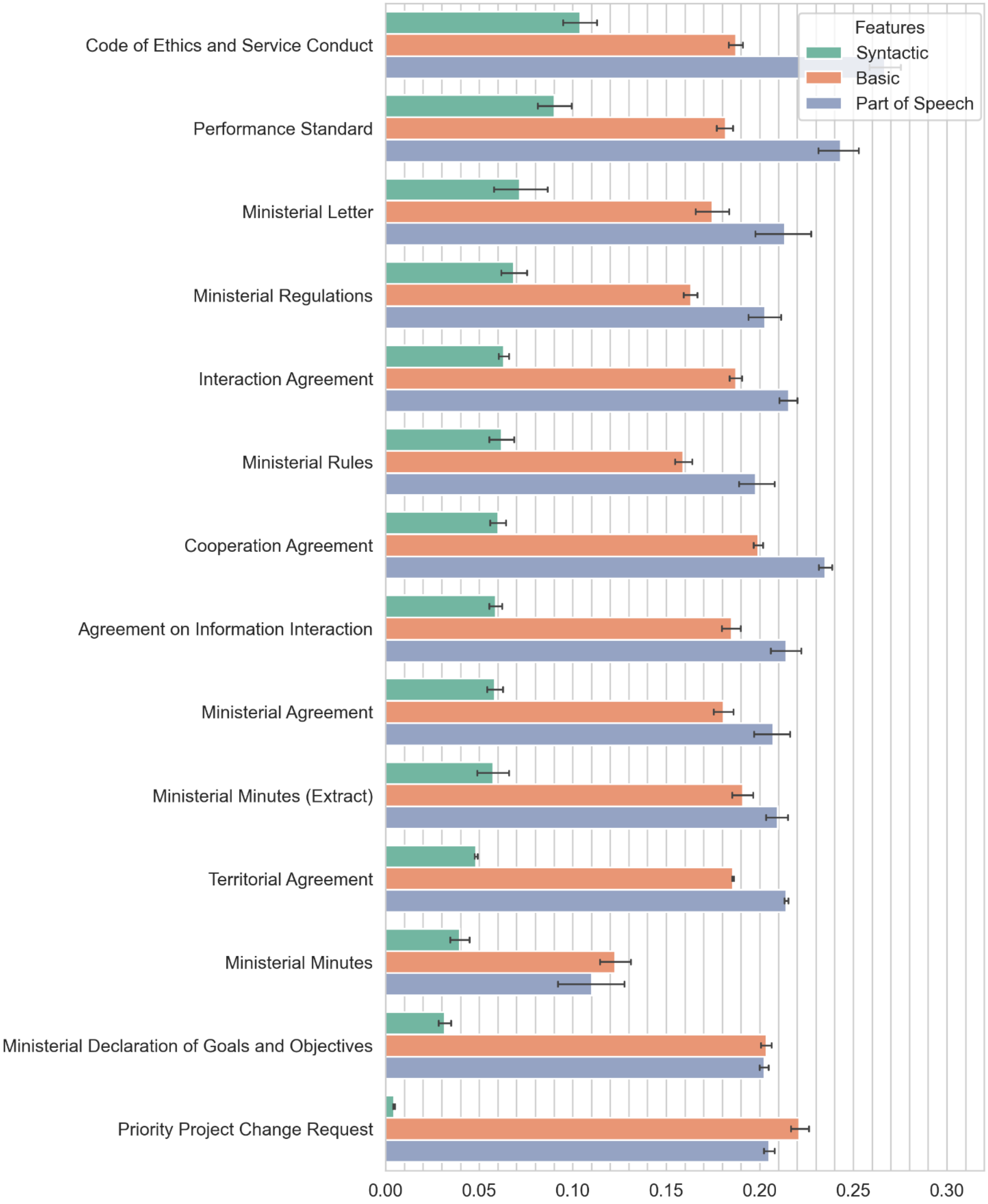


Figure 5. Genres' Complexity within Legislative Sub-style
Рисунок 5. Сложность жанров законодательного подстиля

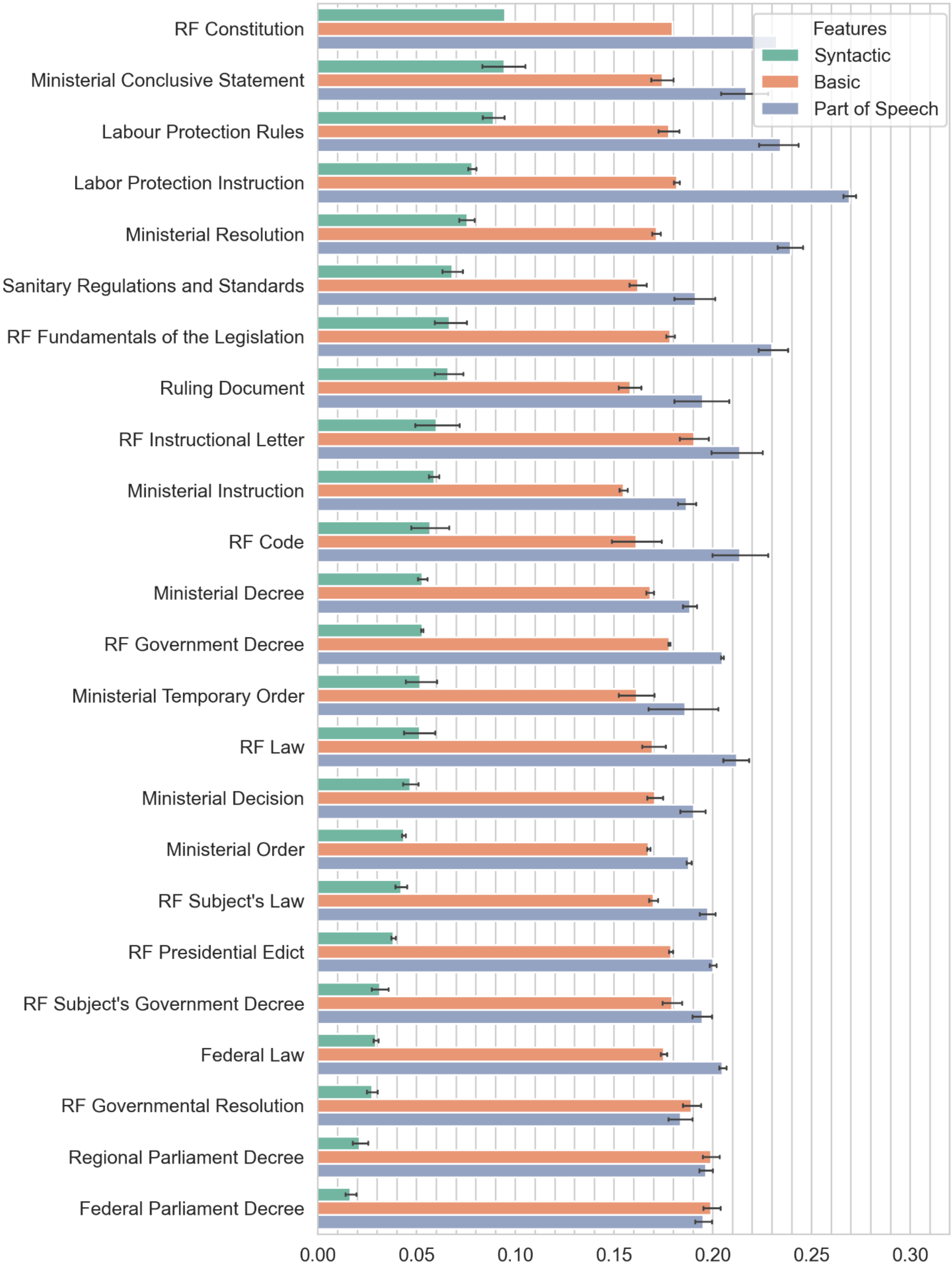
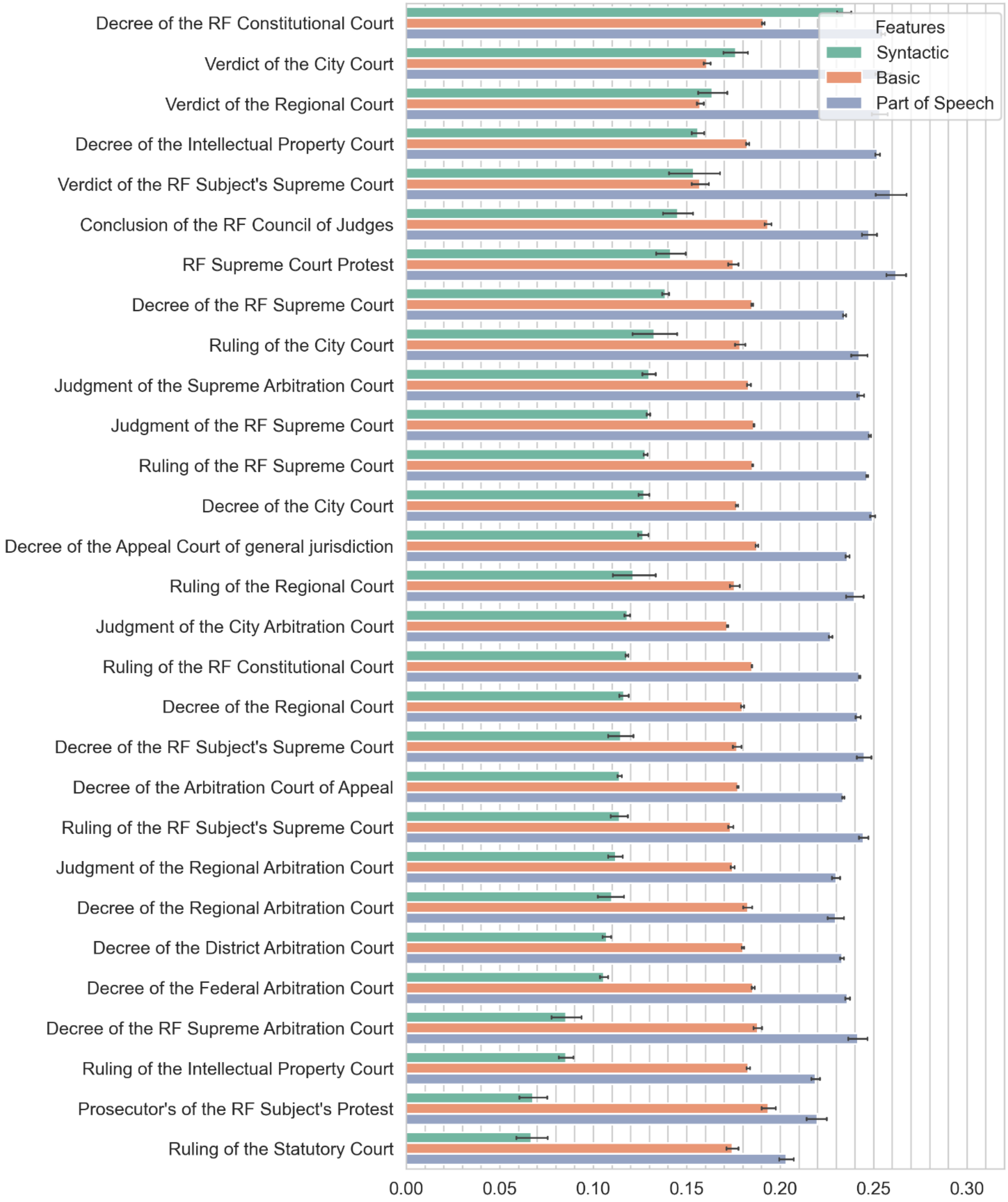


Figure 6. Genres' Complexity within Justiciary Sub-style
Рисунок 6. Сложность жанров юрисдикционного подстиля



Let us give brief 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 (we mean 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");

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 our case, an important component of the complexity model is the consideration of features based on UDPipe markup (Straka and Straková, 2019). Additionally, we used pymorphy2 for part-of-speech tagging and morphological annotation (Korobov, 2015).

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.

Conclusion

This paper 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 paper, we analyze the complexity predictions of the fine-tuned ruBERT model, 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%

[6] https://mintrud.gov.ru/ministry/programms/anticorruption/9/3

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, we interpreted the average values of all syntactic metrics (there are a total of 22 such metrics in our model, see Appendix to this paper). 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 **legislative sub-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.

References

Bhatia, V. K. (1983). *An applied discourse analysis of English legislative writing*, University of Aston in Birmingham, Birmingham, UK. *(In English)*

Bhatia, V. K. (2013). *Analysing Genre: Language use in Professional Settings*, Applied linguistics and language study, Routledge, Taylor & Francis, UK. https://doi.org/10.4324/9781315844992 *(In English)*

Blinova, O. and Tarasov, N. (2022). A hybrid model of complexity estimation: Evidence from Russian legal texts, *Frontiers in Artificial Intelligence,* 5. https://doi.org/10.3389/frai.2022.1008530 *(In English)*

Borisov, A. B. (2010). *Bol'shoj yuridichesky slovar'* [Large legal dictionary], Knizhnyj mir, Moscow, Russia. *(In Russian)*

Dell'Orletta, F., Venturi, G. and Montemagni, S. (2012). Genre-oriented Readability Assessment: a Case Study, *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, The COLING 2012 Organizing Committee, Mumbai, India, 91–98. *(In English)*

Dmitrieva, A. V. (2017). "The art of legal writing": A quantitative analysis of Russian Constitutional Court rulings, *Sravnitel'noe konstitutsionnoe obozrenie*, 118, 125–133. https://doi.org/10.21128/1812-7126-2017-3-125-133 *(In Russian)*

Dodonov, V., Krylova, M., Panov, V., Palatkin, A., Trofimov, V. and Ermakov, V. (2001). *Bol'shoj yuridichesky slovar* [Large legal dictionary], Nauchno-izdatelsky tsentr INFRA-M, Moscow, Russia. *(In Russian)*

Durant, A. and Leung, J. H. (2016). Legal Genres, in *Language and Law: A Resource Book* *for Students, Routledge English Language Introductions*, Routledge, Taylor & Francis, UK, 11–15. https://doi.org/10.4324/9781315436258 *(In English)*

Goźdź-Roszkowski, S. (2007). Legal terms in context: phraseological variation across genres, in *Evidence-Based LSP: Translation, Text and Terminology, Linguistic Insights: Studies in Language and Communication*, Peter Lang AG, Bern, Germany, 455-470. *(In English)*

Goźdź-Roszkowski, S. (2012). *Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study*, Łódź Studies in Language, 22, Peter Lang Verlag, Berlin, Germany. https://doi.org/10.3726/978-3-653-00659-9 *(In English)*

Howe, P. M. (1990). The problem of the problem question in English for academic legal purposes, *English for Specific Purposes*, 9, 215–236. https://doi.org/10.1016/0889-4906(90)90014-4 *(In English)*

Iedema, R. A. M. (1993). Legal English: Subject Specific Literacy and Genre Theory, *Australian Review of Applied Linguistics*, 16, 86–122. *(In English)*

Knutov, A., Plaksin, S., Grigorieva, N., Sinyatullin, R., Chaplinsky, A. and Uspenskaya, A. (2020). *Slozhnost rossiiskih zakonov. Opyt sintaksicheskogo analiza* [Complexity of Russian Laws. The Experience of Syntactic Analysis], HSE University Publishing House, Moscow, Russia. *(In Russian)*

Korobov, M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages, in Khachay, M. Yu., Konstantinova, N., Panchenko, A., Ignatov, D. and Labunets, V. G. (eds.), *Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science*, Springer International Publishing, 320–332. https://doi.org/10.48550/arXiv.1503.07283 *(In English)*

Kozhina, M. N., Duskaeva, L. R. and Salimovsky, V. A. (2011). *Stilistika russkogo yazyka* [Stylistics of the Russian Language], Flinta, Nauka, Moscow, Russia. *(In Russian)*

Kuchakov, R. and Saveliev, D. (2018). *Slozhnost pravovyh aktov v Rossii. Leksicheskoe i sintaksicheskoe kachestvo tekstov*: analiticheskaya zapiska [The complexity of legal acts in Russia: Lexical and syntactic quality of texts: analytic note], European University at Saint Petersburg, Saint Petersburg, Russia. *(In Russian)*

Kurzon, D. (1985). How Lawyers Tell their Tales: Narrative Aspects of a Lawyer's Brief, *Poetics*, 14, 467–481. *(In English)*

Martínez, E., Mollica, F. and Gibson, E. (2022). Poor writing, not specialized concepts, drives processing difficulty in legal language, *Cognition*, 224, 105070. https://doi.org/10.1016/j.cognition.2022.105070 *(In English)*

Mattila, H. E. S. (2013). *Comparative legal linguistics: language of law, Latin and modern lingua francas*, 2nd ed., Ashgate Publishing, Ltd., Farnham, Surrey, UK. *(In English)*

Orts, M. Á. (2015). Power and Complexity in Legal Genres: Unveiling Insurance Policies and Arbitration Rules, *International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique*, 28, 485–505. https://doi.org/10.1007/s11196-015-9429-6 *(In English)*

Saveliev, D. and Kuchakov, R. (2019). *Resheniya arbitrazhnyh sudov subjektov Rossiiskoy Federatsii: leksicheskoe i sintaksicheskoe kachestvo tekstov: analiticheskaja zapiska* [Decisions of arbitration courts of Russian Federation: lexical and syntactic quality of texts, analytic note], European University at Saint Petersburg, Saint Petersburg, Russia. *(In Russian)*

Saveliev, D. A. (2020). Issledovanie slozhnosti predlozheny, sostavlyayushchih teksty pravovyh aktov organov vlasti Rossiiskoy Federatsii [A study in complexity of sentences constituting Russian Federation legal acts], *Pravo. Zhurnal Vysshey shkoly ekonomiki* [Law. Journal of the Higher School of Economics], 1, 50-74. https://doi.org/10.17323/2072-8166.2020.1.50.74 *(In Russian)*

Schwarzkopf, B. S. (1996). Ofitsialno-delovoy yazyk [Official business language], in Graudina, L. K. and Shiryaev, E. N. (eds.), *Kultura russkoy rechi i effektivnost obshheniya* [Culture of Russian speech and effectiveness of communication], Nauka, Moscow, Russia, 270–281. *(In Russian)*

Solganik, G. Ja. (2003). *Stilistika teksta*: Uchebnoe posobie [Text Stylistics: A tutorial], Flinta, Nauka, Moscow, Russia. *(In Russian)*

Solnyshkina, M. I., Solovyev, V. D., Gafiyatova, E. V. and Martynova, E. V. (2022). Slozhnost teksta kak mezhdistsiplinarnaya problema [Text complexity as an interdisciplinary problem], *Voprosy kognitivnoy lingvistiki* [Issues in Cognitive Linguistics], 1, 18-39.

https://doi.org/10.20916/1812-3228-2022-1-18-39 *(In Russian)*

Straka, M. and Straková, J. (2019). *Universal Dependencies 2.5 Models for UDPipe*. URL: http://hdl.handle.net/11234/1-3131 (Accessed 15 January 2023). *(In English)*

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*, Cambridge University Press, Cambridge, UK. *(In English)*

Tessuto, G. (2012). *Investigating English Legal Genres in Academic and Professional Contexts*, Cambridge Scholars Publishing, UK. *(In English)*

Tiersma, P. M. (1986). The Language of Offer and Acceptance: Speech Acts and the Question of Intent, *California Law Review*, 74, 189–232. https://doi.org/10.2307/3480357 *(In English)*

Trosborg, A. (1991). An analysis of legal speech acts in English Contract Law. “It is hereby performed”, *HERMES – Journal of Language and Communication in Business*, 4, 65-90. https://doi.org/10.7146/hjlcb.v4i6.21456 *(In English)*

Trosborg, A. (1995). Statutes and contracts: An analysis of legal speech acts in the English language of the law, *Journal of Pragmatics*, 23, 31–53. https://doi.org/10.1016/0378-2166(94)00034-C *(In English)*

Venturi, G. (2012). Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach, *IALF Porto*, 138-156. *(In English)*

Wang, W. (2019). Text analysis, in McKinley, J. and Heath, R. (eds.), *The Routledge Handbook of Research Methods in Applied Linguistics*, Routledge, London, UK, 453–463. https://doi.org/10.4324/9780367824471 *(In English)*

Список литературы

Борисов А. Б. Большой юридический словарь. М.: Книжный мир, 2010. 848 с.

Дмитриева А. В. «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации // Сравнительное конституционное обозрение. 2017. Т. 118, № 3. С. 125–133.

Додонов В. и др. Большой юридический словарь. М.: Научно-издательский центр ИНФРА-М, 2001. 780 с.

Кожина М. Н., Дускаева Л. Р., Салимовский В. А. Стилистика русского языка. 4-е издание. Москва: Флинта, Наука, 2011. 464 с.

Кучаков Р. К., Савельев Д. А. Сложность правовых актов в России: Лексическое и синтаксическое качество текстов / под ред. Д. Скугаревского. СПб.: Институт проблем правоприменения при Европейском университете в Санкт-Петербурге, 2018. 20 с.

Савельев Д. А. Исследование сложности предложений, составляющих тексты правовых актов органов власти Российской Федерации // Право. Журнал Высшей школы экономики. 2020. Т. 1. С. 50–74.

Савельев Д. А., Кучаков Р. К. Решения арбитражных судов субъектов Российской Федерации: лексическое и синтаксическое качество текстов: аналитическая записка / под ред. Д. Скугаревского. СПб: Институт проблем правоприменения при Европейском университете в Санкт-Петербурге, 2019. 20 с.

Сложность российских законов. Опыт синтаксического анализа / Кнутов А. В., Плаксин С. М., Григорьева Н. Л., Синятуллин Р. Х., Чаплинский А. В., Успенская А. М. М.: Издательский дом НИУ ВШЭ, 2020. 311 с.

Солганик Г. Я. Стилистика текста: Учебное пособие. М.: Флинта; Наука, 2000. 253 с.

Солнышкина М. И., Соловьев В. Д., Гафиятова Э. В., Мартынова Е. В. Сложность текста как междисциплинарная проблема // Вопросы когнитивной лингвистики. 2022. Вып. 1. С. 18-39.

Шварцкопф Б. С. Официально-деловой язык // Культура русской речи и эффективность общения / под ред. Л. К. Граудиной, Е. Н. Ширяева. М.: Наука, 1996. С. 270–281.

Bhatia V. K. An applied discourse analysis of English legislative writing. Birmingham: University of Aston in Birmingham, 1983. 145 p.

Bhatia V. K. Analysing Genre: Language use in Professional Settings. Applied linguistics and language study. London: Routledge, Taylor & Francis, 2013. 264 p.

Blinova O., Tarasov N. A hybrid model of complexity estimation: Evidence from Russian legal texts. Frontiers in Artificial Intelligence. 2022. Vol. 5.

Dell’Orletta F., Venturi G., Montemagni S. Genre-oriented Readability Assessment: a Case

Study // Proceedings of the Workshop on Speech and Language Processing Tools in Education. The COLING 2012 Organizing Committee, Mumbai, 2012. P. 91–98.

Durant A., Leung J. H. Legal Genres // Language and Law: A Resource Book for Students, Routledge English Language Introductions. London: Routledge, Taylor & Francis, 2016. P. 11–15.

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. Bern: Peter Lang AG, 2007. P. 455–470.

Goźdź-Roszkowski S. Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study // Łódź Studies in Language 22. Berlin, Peter Lang Verlag: 2012. 280 p.

Howe P. M. The problem of the problem question in English for academic legal purposes // English for Specific Purposes. 1990. № 9. P. 215–236.

Iedema R. A. M. Legal English: Subject Specific Literacy and Genre Theory // Australian Review of Applied Linguistics. 1993. № 16. P. 86–122.

Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Khachay M. Yu., Konstantinova N., Panchenko A., Ignatov D., Labunets V. G. (ed.), Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science. Springer International Publishing, 2015. P. 320–332.

Kurzon D. How Lawyers Tell their Tales: Narrative Aspects of a Lawyer's Brief // Poetics. 1985. Vol. 14. P. 467–481.

Martínez E., Mollica F., Gibson E. Poor writing, not specialized concepts, drives processing difficulty in legal language // Cognition. 2022. Vol. 224.

Mattila H. E. S. Comparative legal linguistics: language of law, Latin and modern lingua francas, 2nd ed. Farnham, Surrey: Ashgate Publishing, Ltd., 2013. 504 p.

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

Straka M., Straková J. Universal Dependencies 2.5 Models for UDPipe. URL: http://hdl.handle.net/11234/1-3131 (accessed 15.01.2023).

Swales J. M. Genre Analysis: English in Academic and Research Settings. Cambridge, Cambridge University Press: 1990. 274 p.

Tessuto G. Investigating English Legal Genres in Academic and Professional Contexts. Cambridge: Cambridge Scholars Publishing, 2012. 315 p.

Tiersma P. M. The Language of Offer and Acceptance: Speech Acts and the Question of Intent // California Law Review. 1986. Vol. 74. P. 189–232.

Trosborg A. An analysis of legal speech acts in English Contract Law. “It is hereby performed.” // HERMES - Journal of Language and Communication in Business. 1991. Vol. 4. P. 65–90.

Trosborg A. Statutes and contracts: An analysis of legal speech acts in the English language of the law // Journal of Pragmatics. 1995. Vol. 23. P. 31–53.

Venturi G. Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach // IALF Porto. 2012. P. 138–156.

Wang W. Text analysis // McKinley J., Heath R. (ed.), The Routledge Handbook of Research Methods in Applied Linguistics. London: Routledge, 2019. P. 453-463.

Все авторы прочитали и одобрили окончательный вариант рукописи.

All authors have read and approved the final manuscript.

Appendix. Metrics for Assessing Complexity

№	Shorthand	Short Explication
Basic metrics		
1	N_word	Number of tokens (word forms)
2	V_word	Number of types (word forms)
3	N_lemma	Number of tokens (lemmas)
4	V_lemma	Number of types (lemmas)
5	C	Number of characters
6	punct	Number of punctuation characters
7	let	Number of letters
8	N	Number of numeric characters
9	syl	Number of syllables
10	sent	Number of sentences
11	word_long	Number of long word forms
12	word_long_pr	Proportion of long word forms
13	lemma_long	Number of long lemmas
14	lemma_long_pr	Proportion of long lemmas
15	comma_pr	Proportion of commas
16	ASL	Average sentence length in words
17	ASS	Average sentence length in syllables
18	ASW	Average word form length in syllables
19	ACW	Average word form length in letters
20	L	Average number of letters per 100 word forms
21	S	Average number of sentences per 100 word forms
22	TTR_word	SimpleTTR (for word forms)
23	TTR_lemma	SimpleTTR (for lemmas)
24	Yule'sK_word	Yule's K (for word forms)
25	Yule'sK_lemma	Yule's K (for lemmas)
26	Yule'sI_word	Yule's I (for word forms)
27	Yule'sI_lemma	Yule's I (for lemmas)
28	hapax1_pr	Proportion of hapax legomena (for lemmas)
29	hapax2_pr	Proportion of hapax dislegomena (for lemmas)
Words of various part-of-speech classes		
35	Func_word_pr	Analyticity index
36	Verb_pr	Verbality index
37	Noun_pr	Substantivity index
38	Adj_pr	Adjectivity index
39	Pron_pr	Pronominality index
40	Autosem_pr	Autosemanticity index
41	Nouns_pr	Index of noun vocabulary
42	NVR	Noun-Verb ratio
43	Cconj_ pr	Proportion of coordinating conjunctions
44	Sconj_pr	Proportion of subordinating conjunctions
45	Adjs_pr	Proportion of short adjectives
46	Prtf_pr	Proportion of full participles
47	Prts_pr	Proportion of short participles
48	Npro_pr	Proportion of pronouns
49	Pred_pr	Proportion of predicatives
50	Grnd_pr	Share of adverbial participles
51	Infn_pr	Proportion of infinitives

52	Numr_pr	Proportion of numerals
53	Prcl_pr	Proportion of particles
54	Prep_pr	Proportion of one-word prepositions
55	Comp_pr	Proportion of comparative forms
Syntactic metrics		
110	Acl_pr	Proportion of clausal modifiers of a noun (adjectival clause)
111	Acl:relcl_pr	Proportion of relative clause modifiers
112	Advcl_pr	Proportion of adverbial clause modifiers
113	Advmod_pr	Proportion of adverbial modifiers
114	Amod_pr	Proportion of adjectival modifiers
115	Appos_pr	Proportion of appositional modifiers
116	Aux:pass_pr	Proportion of passive auxiliary constructions
117	Cc_pr	Proportion of coordinating conjunction
118	Ccomp_pr	Proportion of clausal complements
119	Compound_pr	Proportion of compounds
120	Conj_pr	Proportion of constructions with conjuncts
121	Cop_pr	Proportion of clauses with copula-like elements
122	Csubj_pr	Proportion of constructions with clausal subject
123	Csubj:pass_pr	Proportion of constructions with clausal passive subject
124	Discourse_pr	Proportion of discourse elements
125	Mark_pr	Proportion of units capable of attaching dependent clauses
126	Nsubj_pr	Proportion of clauses with nominal subject
127	Nsubj:pass_pr	Proportion of clauses with passive nominal subject
128	Nummod_pr	Proportion of numeric modifiers
129	Orphan_pr	Proportion of elliptical predicate constructions
130	Parataxis_pr	Proportion of units connected by a paratactic relationship with other units (discourse-like equivalent of coordination)
131	Xcomp_pr	Proportion of clausal complements without their own subjects

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Conflicts of interests: the authors have no conflicts of interest to declare.

Olga V. Blinova, Candidate of Philological Sciences (Ph.D. in Philology), Associate Professor, St. Petersburg State University, Department of General Linguistics, Higher School of Economics National Research University, HSE Campus in St. Petersburg, Department of Philology, St. Petersburg, Russia.

Ольга Владимировна Блинова, кандидат филологических наук, доцент, Санкт-Петербургский государственный университет, кафедра общего языкознания им. Л. А. Вербицкой, Национальный исследовательский университет «Высшая школа экономики», филиал в Санкт-Петербурге, департамент филологии, Санкт-Петербург, Россия.

Nikita A. Tarasov, Postgraduate Student, St. Petersburg State University, Department of Programming Technologies, St. Petersburg, Russia.

Никита Андреевич Тарасов, аспирант, Санкт-Петербургский государственный университет, кафедра технологии программирования, Санкт-Петербург, Россия.