О. В. Блинова

DOI: 10.24412/1811-1629-2022-2-4-13

ОЦЕНКА СЛОЖНОСТИ РУССКИХ ПРАВОВЫХ ТЕКСТОВ: АРХИТЕКТУРА МОДЕЛИ[1]

OLGA V. BLINOVA
ASSESSING COMPLEXITY OF RUSSIAN LEGAL TEXTS: THE MODEL'S ARCHTECTURE

Ольга Владимировна Блинова

Кандидат филологических наук, доцент

▸ o.blinova@spbu.ru, ovblinova@hse.ru

Санкт-Петербургский государственный университет

199034, Санкт-Петербург, Университетская наб., 7–9

Научно-исследовательский институт «Высшая школа экономики»

190068, Санкт-Петербург, ул. Союза Печатников, 16

Olga V. Blinova

St. Petersburg State University

7/9, Universitetskaya nab., St. Petersburg, 199034

National Research University Higher School of Economics

16, ul. Soiuza Pechatnikov, St. Petersburg, 190068

В статье описана основанная на метриках[2] модель оценки сложности русских правовых текстов. Архитектура модели подразумевает использование 130 метрик, разделённых на следующие категории: «базовые метрики», «формулы читабельности», «учёт слов разных частеречных классов», «n-граммы частеречных помет», «частотность лемм», «словообразовательные модели», «отдельные граммемы», «лексические и семантические признаки, неоднословные выражения», «синтаксические признаки», «оценки связности». Две метрики учитывают гипертекстовые связи и наличие неопределённых контекстов.

Модель способна оценивать и структурную, и понятийную, и интертекстуальную сложность, привлекая и традиционно используемые для предсказания сложности неспецифичные метрики, и метрики стилеспецифичные, разработанные с оглядкой на особенности организации официально-деловых текстов.

При подсчёте морфологических и синтаксических признаков модель обращается к слоям разметки, выполненной UDPipe ("ru-syntagrus") и pymorphy2.

Для обеспечения работы модели создан ряд пользовательских словарей, среди которых: список лексических средств текстового дейксиса[3], список графических сокращений (1,5 тыс. единиц), список аббревиатур (2 тыс. единиц), список юридических терминов (10 тыс. единиц), список абстрактных лемм (17 тыс. единиц), список однословных лексических показателей деонтической возможности и необходимости, список конструкций с лёгкими глаголами.

Значения метрик сложности[4] подсчитаны для всех документов корпуса законов CorCodex, корпуса решений конституционного суда CorDec и корпуса локальных актов CorRIDA (всего порядка 8 млн токенов). Размеченные юридические корпусы, значения метрик сложности и пользовательские словари доступны для скачивания с сайта plaindocument.org.

Ключевые слова: русские правовые тексты; модель оценки сложности; языковые метрики; читабельность.

The paper describes the metrics-based model for assessing complexity of Russian legal texts. The architecture of the model implies the use of 130 metrics divided into following categories: "basic metrics", "readability formulas", "words of different part-of-speech classes", "n-grams of part-of-speech tags", "frequency of lemmas", "word-building patterns", "grammes", "lexical and semantic features, multi-word expressions", "syntactic features", "cohesion assessments". Two metrics take into account hypertext links and the presence of vague contexts.

The model is able to evaluate structural, conceptual, and hypertextual complexity, including both non-specific metrics traditionally used to predict complexity and style-specific metrics developed taking into account the peculiarities of official texts.

When evaluating morphological and syntactic features, the model refers to the markup layers performed by UDPipe ("rusyntagrus") and pymorphy2.

To make the model work a number of user dictionaries are involved, including a list of lexical means of text deixis, a list of graphic abbreviations (1,500 units), a list of acronyms (2,000 units), a list of legal terms (10,000 units), a list of abstract lemmas (17,000 units), a list of lexical indicators of deontic possibility and necessity, a list of light verb constructions.

The values of complexity metrics were calculated for all documents of the CorCodex law corpus, the CorDec corpus of Constitutional court decisions, and the CorRIDA corpus of local acts (about 8 million tokens in total). Annotated legal corpora, complexity metrics, and user dictionaries are available for downloading from plaindocument.org.

Keywords: Russian legal texts, complexity assessment model; linguistic metrics; readability.

1. Введение

Настоящая статья посвящена описанию **модели оценки сложности** русских правовых текстов. **Сложность** текста на языке может пониматься как скрытая переменная, значение которой измеримо для любого связного текста. В некотором общем случае при оценке сложности выбираются различные **параметры** (или признаки), значения которых подсчитываются с помощью соответствующих **метрик сложности**.

В представляемой вниманию читателя статье описывается модель оценки сложности, подразумевающая применение **130 метрик**. При разработке метрик учитывалась специфика правовых текстов, а также опыт исследований языковой сложности и опыт стилометрических исследований.

Обращение к юридическим текстам в контексте изучения сложности закономерно. Указания на сложность и производную от неё неудобопонятность правовых текстов можно считать трюизмами. Скажем, в [Mattila 2013] выделяются **особенности lingua legis** безотносительно определённой правовой традиции или языка. Среди таких особенностей — «**сложность предложений**».

«Сложностью предложений» согласно [Там же] может быть оценена, например, через подсчёт количества зависимых клауз. Х. Маттила указывает также на наличие «усложнённых и бесполезных выражений» типа '*на медленной скорости*' вместо '*медленно*', '*по окончании*' вместо '*после*' и т. п. Кроме того, черты сложности, согласно [Там же], — предпочтение существительных глаголам, а также частотность сложных слов (ср. пример автора, нем. *Klageerzwingungsverfahren* 'исполнительное производство по жалобе') и устойчивых неоднословных выражений (ср. пример автора, франц. *contrat de transfert de processus technologique* 'договор на передачу технологического процесса').

[Azuelos-Atias, Ye 2017] также интересует язык права «вообще»; они указывают на «очень сложный» юридический язык, «сложность языковых структур», интригующую лингвистов, на употребимость «длинных, сложных и избыточных предложений» [Там же: 1, 2, 3]. Авторы подчёркивают, что неудобопонятность правовых текстов для неспециалистов обуславливает не только упомянутая «сложность предложений», то есть **структурная сложность**, но и **понятийная сложность** (выражающаяся в употреблении специальной терминологии). Однако «соломинкой, переломившей спину верблюда», является «тот факт, что важнейшие юридические знания остаются неявными в большинстве юридических документов. **Интертекстуальные ссылки** на информацию, которая хорошо известна специалистам в области права, представлены в юридических текстах почти как рутина — слабыми подсказками» [Там же: 3].

Таким образом, стоит говорить об **объективной сложности** правовых текстов, которую можно измерить. Это предполагает использование метрик, оценивающих структурную, понятийную и интертекстуальную сложность. **Целью** настоящей статьи является описание автоматической модели оценки сложности, подразумевающей учёт указанных типов сложности.

2. Опыт изучения сложности правовых текстов[5]

В исследованиях сложности юридических текстов применяются методы квантитативной лингвистики. Дополнением классических подходов, реализованным в небольшом количестве моделей сложности, является фиксация **гипертекстовых связей** внутри изучаемых текстовых

коллекций и/или фиксация **неопределённых выражений**, см. [Блинова, Белов 2020]. Неопределённые выражения допускают множественность трактовок, а наличие таких выражений в юридическом тексте противоречит его идеологии ясности, точности и однозначности, см. об этом [Кузнецов, Соловьев 2019] и др.

В качестве первого примера можно привести модель [Waltl, Matthes 2014]. Авторы измеряли сложность немецких законов и использовали следующие метрики: количество абзацев, количество предложений, количество слов, **«структурная глубина»** (параметр описывает организацию корпуса законов и гипертекстовые ссылки разной «глубины» на положения того же законодательного акта или другого законодательного акта), **количество внутренних и внешних ссылок**; разнообразие словаря, значение стандартной метрики читабельности. [Waltl, Matthes 2014] включили в оценку сложности и **«неопределённость»**, вычисляемую так: оценивается количество вхождений в тексты немецких законов неопределённых (vague) прилагательных типа «*адекватный*», «*разумный*».

Вторым примером может стать модель [Owens, Wedeking 2011]. Указанные авторы исследовали решения Верховного суда США и измеряли том числе степень уверенности / неуверенности пишущего субъекта в утверждении, оцениваемую через количество вхождений слов типа '*maybe*', '*fairly*', '*perhaps*', '*absolutely*', '*clearly*', количество употреблений эпистемических глаголов, выражающих знание, мнение, предположение, типа '*think*', '*know*', '*consider*' и т. п.

Русские юридические тексты также привлекали внимание исследователей сложности. Наиболее простая модель оценки подразумевает применение формул читабельности, см., например, [Солнышкина, Кисельников 2015]. Так, в статье [Дмитриева 2017] корпус текстов решений Конституционного суда РФ исследован с применением адаптированной формулы Флеша-Кинкейда [Оборнева 2005].

Несколько более изощрённые модели подразумевают применение метрик синтаксической сложности. В [Кучаков, Савельев 2018], [Савельев, Кучаков 2019] для оценки сложности авторы использовали одну метрику лексического разнообразия (меру TTR, отношение числа уникальных слов документа (types) ко всем словам документа (tokens), значение меры зависимо от длины текста) и одну синтаксическую метрику (расстояние в словах между главными и зависимыми узлами в предложении, dependency length), вычисляемое так: «для каждого конкретного текста взято одно значение, которое является максимальным для всех предложений текста» [Там же]).

Наконец, в [Кнутов и др. 2020] использованы девять метрик сложности: «доля глаголов в страдательном залоге», «доля глаголов от общего количества слов в тексте», «среднее количество слов в субстантивных именных словосочетаниях», «среднее количество причастных оборотов, расположенных в предложениях после определяемого слова, на одно предложение», «среднее количество деепричастных оборотов на одно предложение», «среднее количество слов в предложениях», «среднее расстояние между зависимыми словами в предложении», «среднее количество грамматических основ (предикативных основ, предикативных ядер) предложения (подлежащее, сказуемое или одно из них) в одном предложении», «среднее количество слов в абзаце».

Таким образом, авторы работ, выполненных на русском материале, сконцентрировались на изучении сложности языка законов и языка судебных решений. Кроме того, для измерения сложности использовались либо только формулы читабельности, либо другие достаточно немногочисленные метрики. Наконец, авторы не проводили тестирования своих моделей оценки сложности.

3. Архитектура модели

В нашей модели для оценки сложности используется 130 метрик, разделённых на следующие категории:

1. «базовые метрики»,
2. «формулы читабельности»,
3. «учёт слов разных частеречных классов»,
4. «n-граммы частеречных помет»,
5. «частотность лемм»,
6. «словообразование»,
7. «отдельные граммемы»,

8. «лексические и семантические признаки, неоднословные выражения»,
9. «синтаксические признаки»,
10. «оценки связности текста».

3.1. Базовые метрики сложности

Модель предусматривает использование 28 базовых метрик. Их можно разделить на **базовые квантитативные** и **базовые лексические**. Первые нацелены в том числе на учёт долей вхождений в тексты длинных слов и длинных предложений. Длинными словами считаются слова, состоящие из 4-х и более слогов. Примерами таких метрик могут стать, в частности: ASL — «средняя длина предложения в словах», ASW — «средняя длина словоформы в слогах», S — «среднее число предложений на 100 словоформ» и пр.

Базовые лексические метрики подразумевают подсчёт индексов лексического разнообразия (простого TTR для словоформ и лемм; производных от TTR метрик Yule's K и Yule's I, значение которых не зависит от длины текста, см. об этом [Blinova et al. 2020]), а также подсчёт долей гапаксов.

3.2. Формулы читабельности

Использование формул — общераспространённый метод оценки сложности. Сейчас он применяется в комбинации с другими, более изощрёнными, методами, подробнее см., например, [Benjamin 2012], и встраивается в разнообразные текстометрические ресурсы.

В модели используется пять формул: адаптированная формула Флеша-Кинкейда [Solnyshkina et al. 2018], адаптированная формула SMOG, адаптированная формула подсчёта автоматизированного индекса читабельности ARI, индекс Дейла-Чейл, индекс Колман-Лиау, см. [Бегтин 2016].

3.3. Учёт слов разных частеречных классов

В текстах официально-делового стиля при сравнении с другими стилями меняется частота вхождения слов разных частей речи. Сказанное подтверждается, например, в [Браславский 2001], где автор наблюдает «монотонный рост средних долей существительных и прилагательных и монотонное же уменьшение долей местоимений, наречий, глаголов и частиц от разговорного к официально-деловому стилю», см. кроме того, [Поспелова, Ягунова 2014], [Клышинский и др. 2013] и др.

В работе [Дружкин 2016] выделен набор признаков, показавших **высокую положительную корреляцию со сложностью текстов**, среди них есть и частеречные:

- полная форма прилагательного и причастия,
- форма причастия,
- прилагательные и существительные.

В литературе подчёркивается свойственный текстам официально-делового стиля и сложным текстам **рост доли существительного и падение доли глагола в личной форме**.

В нашей модели 22 метрики, учитывающие доли вхождений слов разных частей речи, разработаны с учётом различий между использованными инструментами разметки — UDPipe (лемматизация, частеречная разметка, синтаксическая разметка) и pymorphy2 (второй слой частеречной разметки, подробный морфологический анализ), то есть различий между наборами помет частеречной разметки [Straka, Straková 2019], [Korobov 2015].

Вслед за [Журавлёв 1988] в модель введены: индекс аналитичности (отношение числа служебных слов к общему числу слов в тексте); индекс глагольности; индекс субстантивности; индекс адъективности; индекс местоименности; индекс автосемантичности (отношение числа значащих слов к общему числу слов; «незначащими» считаются все служебные слова и местоимения).

Метрика Comp_pr обращается к слою частеречной разметки, но введена с прицелом на оценку доли вхождений прилагательных и наречий в форме сравнительной степени (то есть градуируемых прилагательных и наречий, потенциально участвующих в формировании неопределённых (vague) контекстов, см. об этом раздел 2 выше, а также [Блинова, Белов 2020]).

Полный список метрик дан на сайте проекта РНФ *plaindocument.org/corpora*.

3.4. N-граммы частеречных помет

Информацию о встречаемости n-грамм частеречных помет решено привлечь для анализа сложности также под влиянием литературы

о квантитативном анализе стиля, в частности, работ [Клышинский и др. 2013], [Антонова и др. 2011].

[Там же] предложена так называемая «**формула динамичности / статичности**», призванная отделить тексты, в которых описывается множество событий, («динамические тексты») от текстов «статических». Эта метрика хорошо противопоставляет тексты официально-делового стиля другим текстам (официально-деловые тексты более «статичны»).

В описываемой в настоящей статье модели используется 13 метрик указанной категории. Специально стоит прокомментировать биграммы вида 'NOUN + NOUN', триграммы вида 'NOUN + NOUN + NOUN' и биграммы вида 'NOUN + NOUN,*gent'. Их использование нацелено в том числе на выделение именных групп с генитивными аргументами.

В дескриптивной литературе по стилистике описываются, а в прескриптивной литературе — строго осуждаются так называемые «цепочки форм родительного падежа» (ср. также понятия «нанизывания падежей», «родительный приимённый»). Исследование именных групп с более чем двумя генитивными аргументами на материале корпуса CorCodex показало, что наибольшая наблюдаемая длина группы с ветвящимся генитивом достигает восьми элементов, см. [Веденина 2021]. Конструкции с несколькими генитивными аргументами в литературе по стилистике эксплицитно оцениваются как трудные для восприятия, ср., например, цитату из [Голуб 2001]: «Затрудняет восприятие текста нанизывание одинаковых грамматических форм, которые последовательно зависят друг от друга <…>. Эпифора часто возникает при нанизывании форм родительного падежа, что обычно связано с влиянием официально-делового стиля».

3.5. Частотность лемм

При оценке сложности принято учитывать длину слов текста и их «знакомость» читающему. Длина слов текста в нашей модели учитывается в базовых квантитативных метриках и в некоторых формулах читабельности. «Знакомость» операционализируется через применение информации об общеязыковой частотности лемм текста.

Исследования на материале русского языка подтверждают, что признаки, включающие информацию о частотности слов и / или их включённости в лексические минимумы, хорошо предсказывают сложность, см. [Ivanov et al. 2018] а также [Sharoff et al. 2008], [Solovyev et al. 2018].

В рамках описываемой модели для аккуратного учёта данных о частотности лемм на базе больших русских корпусов создан частотный список, в котором с применением меры Ципфа (Zipf value) все леммы (примерно 1 млн) распределены по 9-ти частотным диапазонам, см. о методе [Blinova, Tarasov et al. 2020]. Соответственно, наша модель оценки сложности способна учитывать доли лемм, принадлежащих каждой из частотных зон и различать высокочастотные, среднечастотные и низкочастотные леммы.

Итоговое значение Zipf value в сводном частотном списке принимает значения от 0 (наиболее низкочастотные леммы) до 8 (высокочастотные леммы). При оценке сложности учитываются доли лемм каждого из девяти частотных диапазонов (оцениваемые в 9-ти метриках описываемой категории).

3.6. Словообразование

Для диагностики сложности в [Дружкин 2016] использовались концовки («хвосты») словоформ и лемм. Высокую положительную корреляцию со сложностью показали, в частности, трёхбуквенные «хвосты» *ние, сть, ция, вие, тво, щий, кий, тья*, четырёхбуквенные «хвосты» *ение, ание, ьный, ость, нный, ация, ский, твие, ящий, ство* и пр. [Там же]. Таким образом, практика изучения сложности показывает, что для оценки сложности текста разумно использовать и словообразовательную информацию.

Производные слова, образованные с помощью аффиксов, длиннее производящих. Кроме того, они сложнее морфологически, так как состоят из большего количества морфем. Указанное свойство делает единицы более перцептивно трудными, что подтверждается экспериментально. Например, в [Нагель 2017: 18] читаем: «именно структурная композицио-

нальность, а не частотность единиц осложняет когнитивную обработку синкретичного производного в сравнении с непроизводным словом, что проявляется в замедлении времени реакции у информантов, выполняющих задание на принятие лексического решения»; подробнее об обработке производных слов см. [Слюсарь 2018].

В нашей модели словообразовательная информация извлекается с уровня лемм, в каждом документе подсчитываются доли лемм вида **ция*, **ние*, **вие*, **тие*, **ист*, **изм*, **ура*, **ище*, **ство*, **ость*, **овка*, **атор*, **итор*, **тель*, **льный*, **овать* (то есть учитываются вхождения в тексты целого ряда отглагольных, отадъективных и некоторых других производных существительных, а также избранных отглагольных прилагательных и производных глаголов).

3.7. Отдельные граммемы

При анализе явлений грамматического уровня предлагается, в частности, идентифицировать в текстах редкие и/или сложные морфологические явления [Collins-Thompson 2014]. Для русского языка эта проблематика разработана достаточно мало.

В модели используется 17 метрик, учитывающих, в частности: долю словоформ в родительном, творительном, дательном падеже, долю существительных среднего рода, долю глаголов в форме 3-го лица, долю полных страдательных причастий, долю кратких страдательных причастий и др.

Для того, чтобы подробно мотивировать выбор каждой метрики, потребуется отдельная статья. Приведу только один пример. Творительный падеж кодирует агенса в пассивных конструкциях. Известно, что в официально-деловых текстах пассивные конструкции частотны (утверждение справедливо и для правовых текстов на языках, отличных от русского). [Charrow, Charrow 1979] в работе о восприятии правовых текстов показали, что пассивные конструкции труднее активных, и что пассивные конструкции без выраженного агенса (“truncated passive”), вопреки некоторым теоретическим ожиданиям, легче полных.

3.8. Лексические и семантические признаки, неоднословные выражения

В 11-ти метриках рассматриваемой категории также учтены черты текстов официально-делового стиля. Объёдиняет метрики и необходимость применения пользовательских словарей, в которых в виде списков перечислены единицы, вхождения которых учитывает модель.

Среди специально разработанных пользовательских словарей:

- список лексических средств текстового дейксиса типа *(выше/ниже)названный*, *(выше/ниже)перечисленный*, *данный* и пр.
- список графических сокращений (1,5 тыс. единиц) и аббревиатур (2 тыс. единиц),
- список юридических терминов (10 тыс. единиц),
- список абстрактных лемм (17 тыс. единиц),
- список однословных лексических показателей деонтической возможности и необходимости, ср.: *дозволить*, *должен*, *допустимо*, *запрещать*, *можно*, *надлежащий*, *неправомерно* и т. д.,
- список конструкций с лёгкими глаголами (учтено 6 тыс. последовательностей лемм типа ‘*оказывать содействие*’, ‘*давать оценка*’, ‘*осуществлять подготовка*’).

Список юридических терминов собран на основе обширных юридических словарей [Борисов 2010], [Додонов и др. 2001], он включает и однословные, и неоднословные термины.

Вслед за некоторыми исследователями, см. [Солнышкина, Кисельников 2015], [Solovyev et al. 2020], при оценке сложности решено учитывать и долю абстрактных лемм.

Список абстрактных лемм длиной 17075 единиц включает, к примеру, такие лексемы, как *‘агрессия’*, *‘аффект’*, *‘закон’*, *‘право’*, *‘оправдание’*, *‘обвинение’*, *‘узаконивание’*, *‘беззаконность’*, *‘законность’*, *‘законопослушность’*, *‘незаконность’*, *‘противозаконность’*, *‘правонарушение’*, *‘преступление’* и пр.

В заключение раздела важно заметить, что по крайней мере некоторые метрики обсуждаемой категории призваны оценивать **понятийную сложность** анализируемых текстов. Сказанное ка-

сается и собственно юридических терминов, и абстрактных слов, и аббревиатур.

Для учёта ссылок на федеральные законы РФ в состав метрик описываемой категории введена метрика «FZ_pr» (доля указаний на федеральные законы). Она описывает гипертекстовые связи, явление, названное в [Waltl, Matthes 2014] «структурной глубиной». Соответственно, метрика служит для оценки не собственно языковой сложности, но сложности организации корпуса правовых текстов (если корпус понимать не столько лингвистически, сколько юридически).

3.9. Синтаксические признаки

Возможности анализа синтаксической сложности обуславливаются и ограничиваются форматом синтаксической разметки, возможностями парсера. В представляемой модели для разметки использован анализатор UDPipe (модель "ru-syntagrus", см. [Straka, Straková 2019]).

Наша модель использует 21 синтаксическую метрику и учитывает, к примеру: клаузальные модификаторы имени, в том числе относительные клаузы; сентенциальные обстоятельства; конструкции с сентенциальными дополнениями; аппозитивные конструкции и мн. др. (заинтересованный читатель может обратиться к странице сайта *plaindocument.org*).

3.10. Оценки связности текста

Для оценки связности в состав метрик введена мера «Cohes_1» (количество повторов существительных в соседних предложениях). Кроме того, использована метрика «Cohes_2», учитывающая количество повторов граммем времени и вида у глаголов в личной форме (в соседних предложениях). «Cohes_2» призвана охарактеризовать однотипность / неоднотипность глагольных форм, см. об этом у [Голуб 2001]: «в тексте <официально-делового стиля — О. Б.> обычно повторяются однотипные конструкции: <…>. Такое построение высказываний не только способствует предельной ясности формулировок, но и служит достижению единообразия в изложении».

Специфика организации когезии (структурной связности) в юридических текстах учтена и в одной из лексических метрик, учитывающей вхождения средств текстового дейксиса (см. радел 3.8). Кроме того, в категории метрик, основанных на синтаксическом слое разметки UDPipe, присутствует признак "Discourse" (соответствующая метрика подсчитывает лексические средства, организующие членение и связность).

4. Заключение

Описанная модель оценки сложности русских правовых текстов подразумевает использование 130 метрик, обращающихся к лексике, семантике, синтаксису и связности текста, частично учитывающих сочетаемость и некоторые словообразовательные модели. Кроме того, добавлена метрика, учитывающая гипертекстовые связи (что особенно важно при рассмотрении корпуса законов), а также метрика, способная диагностировать неопределённые контексты.

Модель учитывает и структурную, и понятийную, и (в некоторой степени) интертекстуальную сложность; включает и традиционно используемые для предсказания сложности неспецифичные метрики, и метрики стилеспецифичные, разработанные с оглядкой на особенности организации официально-деловых текстов.

Значения 130 метрик сложности подсчитаны для всех документов корпуса законов CorCodex, корпуса решений конституционного суда CorDec и корпуса локальных актов CorRIDA (всего порядка 8 млн токенов). Сами размеченные корпусы и значения метрик сложности можно скачать с сайта проекта *plaindocument.org*.

Проведено тестирование модели на внешних наборах данных (на наборе данных "plainrussian" [Бегтин 2016] и на наборе русских школьных учебников [Solovyev, Solnyshkina et al. 2019]. В ходе тестирования выявлены метрики, сработавшие в задаче классификации по сложности наилучшим образом. Подробные результаты тестирования предстоит опубликовать.

ПРИМЕЧАНИЯ

[1]Разделы 1 и 3 подготовлены при поддержке гранта РНФ № 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика».

[2]Метрики разрабатываются для измерения представленности в текстах языковых признаков. Например, «слово длиннее 4-х слогов» — это языковой признак. Доля длинных слов в тексте (отношение числа длинных слов ко всем словам текста) — это метрика.

[3]При текстовом дейксисе отсылки производятся «в пространстве текста», точкой отсчёта является место текста, в котором «находятся» адресат и адресант, а дейктическая единица отсылает не к элементу внеязыковой действительности, а к другому (предшествующему или последующему) элементу текста, ср. *вышеуказанный*.

[4]Значения метрик сложности позволяют сравнивать тексты по сложности (или предсказывать трудность текстов для чтения). Сильная представленность признака (высокое значение метрики) не обязательно свидетельствует о большей сложности. Например, чем больше в текстах длинных слов, тем они сложнее; однако чем больше в текстах глаголов в личной форме, тем в некотором общем случае они проще (в случае, если падение доли глагола сопровождается возрастанием доли существительного).

[5]Раздел 2 написан в рамках выполнения государственного задания СПбГУ по проекту НИИ Проблем государственного языка.

ЛИТЕРАТУРА

Антонова и др. 2011 — Антонова А. Ю., Клышинский Э. С., Ягунова Е. В. Определение стилевых и жанровых характеристик коллекций текстов на основе частеречной сочетаемости. В сб.: *Труды международной конференции «Корпусная лингвистика–2011»*. СПб., 2011. С. 80–85.

Бегтин 2016 — Бегтин И. В. *Оценка читабельности текста*. URL: https://github.com/ivbeg/readability.io/wiki/API (дата обращения: 01.02.2022).

Блинова, Белов 2020 — Блинова О. В., Белов, С. А. Языковая неоднозначность и неопределённость в русских правовых текстах. *Вестник Санкт-Петербургского университета. Право*. 2020, (11 (4)): 774–812.

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

Браславский 2001 — Браславский П. Морфологический строй функциональных стилей (на материале документов Internet). *Известия Уральского государственного университета.* 2001, 21: 9–7.

Веденина 2021 — Веденина У. А. *Цепочки зависимых существительных в современных русских юридических документах: выпускная квалификационная работа бакалавра*. СПб., 2021.

Голуб 2001 — Голуб И. Б. *Стилистика русского языка*. 3-е изд., испр. М.: Рольф, 2001. 240 с.

Дмитриева 2017 — Дмитриева А. В. «Искусство юридического письма»: количественный анализ решений Конституционного суда Российской Федерации. *Сравнительное конституционное обозрение*. 2017, (118 (3)): 125–133.

Додонов и др. 2001 — Додонов В. Н., Ермаков В. Д., Крылова М. А. *Большой юридический словарь*. М.: ИНФРА-М, 2001. 780 с.

Дружкин 2016 — Дружкин К. Ю. *Метрики удобочитаемости для русского языка: выпускная квалификационная работа магистра*. М., 2016.

Журавлёв 1988 — Журавлёв А. Ф. Опыт квантитативно-типологического исследования разновидностей устной речи. В сб.: *Разновидности городской устной речи*. М.: Наука, 1988. С. 84–150.

Клышинский и др. 2013 — Клышинский Э. С., Кочеткова Н. А., Мансурова О. Ю., Ягунова Е. В., Максимов В. Ю., Карпик О. В. *Формирование модели сочетаемости слов русского языка и исследование ее свойств*. Препринты ИПМ им. М. В. Келдыша. 2013, 41. URL: https://keldysh.ru/papers/2013/prep2013_41.pdf (дата обращения: 01.02.2022).

Кнутов и др. 2020 — Кнутов А. В., Плаксин С. М., Григорьева Н. Л., Синятуллин Р. Х., Чаплинский А. В., Успенская А. М. *Сложность российских законов. Опыт синтаксического анализа.* М.: Изд. дом Высшей школы экономики, 2020. 311 с.

Костенко 2005 — Костенко М. А. Правовая лингвистика в законотворческом процессе. *Известия ЮФУ. Технические науки.* 2005, (9 (53)): 127–132.

Кузнецов, Соловьев 2019 — Кузнецов С. А., Соловьев А. А. Конституция Российской Федерации в аспекте требований к русскому языку как государственному. *Вестник Волгоградского государственного университета. Серия 2, Языкознание.* 2019, (18 (2)): 27–36.

Кучаков, Савельев 2018 — Кучаков Р. К., Савельев Д. А. Сложность правовых актов в России: Лексическое и синтаксическое качество текстов. СПб.: ИПП ЕУСПб, 2018. 20 с.

Нагель 2017 — Нагель О. В. *Словообразовательные механизмы в процессах восприятия, идентификации и использования языка. автореф. дисс. … докт. филол. наук.* Томск, 2017. 45 с.

Оборнева 2005 — Оборнева И. В. Автоматизация оценки качества восприятия текста. *Вестник Московского городского педагогического университета*. 2005, (2): 221–233.

Поспелова, Ягунова 2014 — Поспелова А. Г., Ягунова Е. В. Опыт применения стилевых и жанровых характеристик для описания стилевых особенностей коллекций текстов. *Новые информационные технологии в автоматизированных системах.* 2014, (17): 347–356.

Савельев, Кучаков 2019 — Савельев Д. А., Кучаков Р. К. *Решения арбитражных судов субъектов Российской Федерации: лексическое и синтаксическое качество текстов: аналитическая записка*. СПб: ИПП ЕУСПб, 2019. 20 с.

Слюсарь 2018 — Слюсарь Н. А. *Экспериментальное исследование ряда основных понятий теоретической морфологии (на материале русского языка): регулярность, синкретизм, маркированность. Резюме дисс. … докт. филол. наук.* М., 2018. 48 с.

Солнышкина, Кисельников 2015 — Солнышкина М. И., Кисельников А. С. Сложность текста: этапы изучения в отечественном прикладном языкознании. *Вестник Томского государственного университета. Филология.* 2015, (6 (38)): 86–99.

Azuelos-Atias, Ye 2017 — Azuelos-Atias S., Ye N. On drafting, interpreting, and translating legal texts across languages and cultures. *International Journal of Legal Discourse.* 2017, (2 (1)): 1–12.

Benjamin 2012 — Benjamin R. G. Reconstructing readability: recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*. 2012, (24 (1)): 63–88.

Blinova et al. 2020 — Blinova O., Belov S., Revazov M. Decisions of Russian Constitutional Court: Lexical Complexity Analysis in Shallow Diachrony. In: *CEUR Workshop Proceedings (Proceedings of the International Conference «Internet and Modern Society» IMS-2020)*. ITMO University, St. Petersburg, Russia. P. 61–74.

Blinova, Tarasov et al. 2020 — Blinova O. V., Tarasov N. A., Modina V. V., Blekanov I. S. Modeling Lemma Frequency Bands for Lexical Complexity Assessment of Russian Texts. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020" (Moscow, June 17–20, 2020)*. 2020, (19 (26)): 76–92.

Charrow, Charrow 1979 — Charrow R. P., Charrow V. R. Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions. *Columbia Law Review*. 1979, (79 (7)): 1306–1374.

Collins-Thompson 2014 — Collins-Thompson K. Computational assessment of text readability: a survey of current and future research. Recent Advances in Automatic Readability Assessment and Text Simplification. *Special issue of International Journal of Applied Linguistics*. 2014, (165 (2)): 97–135.

Ivanov et al. 2018 — Ivanov V. V., Solnyshkina M. I., Solovyev V. D. Efficiency of text readability features in Russian academic texts. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*. 2018, (17): 277–287.

Korobov 2015 — Korobov M. *Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts*. 320–332.

Mattila 2013 — Mattila H. E. S. *Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas*. Routledge, 2013. 504 p.

Owens, Wedeking 2011 — Owens R. J., Wedeking J. P. Justices and legal clarity: Analyzing the complexity of US Supreme Court opinions. *Law & Society Review*. 2011, (45 (4)): 1027–1061.

Solnyshkina et. al 2018 — Solnyshkina M., Ivanov V., Solovyev V. Readability Formula for Russian Texts: A Modified Version. In: *Proceedings of the 17th Mexican International Conference on Artificial Intelligence, MICAI 2018*. Guadalajara, Mexico, part II. P. 132–145.

Solovyev et al. 2018 — Solovyev V., Ivanov V., & Solnyshkina M. Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics. *Journal of Intelligent & Fuzzy Systems*. 2018, (34 (5)): 3049–3058.

Solovyev et al. 2020 — Solovyev V. D., Solnyshkina M. I., Andreeva M., Danilov A., Zamaletdinov R. Text Complexity and Abstractness: Tools for the Russian Language. IMS. In: *CEUR Workshop Proceedings (Proceedings of the International Conference «Internet and Modern Society» IMS-2020)*. ITMO University, St. Petersburg, Russia, 2020. P. 75–87.

Solovyev, Solnyshkina et al. 2019 — Solovyev V., Solnyshkina M., Ivanov V., Batyrshin I. Prediction of reading difficulty in Russian academic texts. *Journal of Intelligent & Fuzzy Systems*. 2019, (36 (5)): 4553–4563.

Straka, Straková 2019 — Straka M., Straková J. Universal Dependencies 2.5 Models for UDPipe (2019-12-06). In: *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2019*. URL: http://hdl.handle.net/11234/1-3131 (date of access: 01.02.2022).

Tiersma 1999 — Tiersma Peter M. *Legal Language*. Chicago, London: University of Chicago Press, 1999. 314 p.

Waltl, Matthes 2014 — Waltl B., Matthes F. *Towards Measures of Complexity: Applying Structural and Linguistic Metrics to German Laws*. JURIX. 2014: 153–162.

Wydick, Sloan 2019 — Wydick R. C., Sloan A. E. *Plain English for lawyers*. Sixth edition. Durham, North Carolina: Carolina Academic Press, LLC, 2019. 178 p.

REFERENCES

Антонова и др. 2011 — Antonova A. Iu., Klyshinskii E. S., Iagunova E. V. Determination of stylistic and genre characteristics of text collections based on part-of-speech compatibility. In: *Trudy mezhdunarodnoi konferentsii «Korpusnaia lingvistika–2011»*. Saint Petersburg, 2011. P. 80–85. (In Russian)

Бегтин 2016 — Begtin I. V. *Readability assessment*. URL: https://github.com/ivbeg/readability.io/wiki/API (date of access: 01.02.2022). (In Russian)

Блинова, Белов 2020 — Blinova O. V., Belov S. A. Linguistic ambiguity and vagueness in Russian legal texts. *Vestnik Sankt-Peterburgskogo universiteta. Pravo.* 2020, (11 (4)): 774–812. (In Russian)

Борисов 2010 — Borisov A. B. *Large law dictionary*. Moscow: Knizhnyi mir, 2010. 848 p. (In Russian)

Браславский 2001 — Braslavskii P. Morphological structure of functional styles (based on Internet documents). *Izvestiia Ural'skogo gosudarstvennogo universiteta*. 2001, (21): 9–7. (In Russian)

Веденина 2021 — Vedenina U. A. *Nouns chains in modern Russian legal documents (according to corpus data): bachelor's thesis*. Saint Petersburg, 2021. (In Russian)

Голуб 2001 — Golub I. B. *Stylistics of the Russian language*. 3rd ed. Moscow: Rol'f, 2001. 240 p. (In Russian)

Дмитриева 2017 — Dmitrieva A. V. "The art of legal writing": A quantitative analysis of Russian Constitutional Court rulings. *Sravnitel'noe konstitutsionnoe obozrenie.* 2017, (118 (3)): 125–133. (In Russian)

Додонов и др. 2001 — Dodonov V. N., Ermakov V. D., Krylova M. A. *Large law dictionary*. Moscow: INFRA-M, 2001. 780 p. (In Russian)

Дружкин 2016 — Druzhkin K. Iu. *Readability metrics for Russian: master's theses*. Moscow, 2016. (In Russian)

Журавлёв 1988 — Zhuravlev A. F. Experience of Quantitative-Typological Study of Varieties of Oral Speech. In: *Raznovidnosti gorodskoi ustnoi rechi*. Moscow: Nauka, 1988. P. 84–150. (In Russian)

Клышинский и др. 2013 — Klyshinskii E. S., Kochetkova N. A., Mansurova O. Iu., Iagunova E. V., Maksimov V. Iu., Karpik O. V. *Formation of the Russian Word Combinability Model and Study of its Properties. Preprinty IPM im. M. V. Keldysha*. 2013, 41. URL: https://keldysh.ru/papers/2013/prep2013_41.pdf (дата обращения: 01.02.2022). (In Russian)

Кнутов и др. 2020 — Knutov A. V., Plaksin S. M., Grigor'eva N. L., Siniatullin R. Kh., Chaplinskii A. V., Uspenskaia A. M. *The Complexity of Russian Laws. The Experience of Syntactic Analysis.* Moscow: Izd. dom Vysshei shkoly ekonomiki, 2020. 311 p. (In Russian)

Костенко 2005 — Kostenko M. A. Legal linguistics in the legislative process. *Izvestiia IuFU. Tekhnicheskie nauki.* 2005, (9 (53)): 127–132. (In Russian)

Кузнецов, Соловьев 2019 — Kuznetsov C. A., Solov'ev A. A. The Constitution of the Russian Federation in the Aspect of Requirements for Russian as a State Language. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriia 2, Iazykoznanie*. 2019, (18 (2)): 27–36. (In Russian)

Кучаков, Савельев 2018 — Kuchakov R. K., Savel'ev D. A. *The complexity of legal acts in Russia: Lexical and syntactic quality of texts: analytic note.* Saint Petersburg: ИПП ЕУСПб, 2018. 20 p. (In Russian)

Нагель 2017 — Nagel' O. V. *Word-formation mechanisms in the processes of perception, identification, and use of language: Author's abstract of the Doctor's Thesis.* Tomsk, 2017. 45 p. (In Russian)

Оборнева 2005 — Oborneva I. V. Automation of text perception quality assessment. *Vestnik Moskovskogo gorodskogo pedagogicheskogo universiteta.* 2005, (2): 221–233. (In Russian)

Поспелова, Ягунова 2014 — Pospelova A. G., Iagunova E. V. Experience using style and genre characteristics to describe the stylistic features of texts collections. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh.* 2014, (17): 347–356. (In Russian)

Савельев, Кучаков 2019 — Savel'ev D. A., Kuchakov R. K. *Decisions of arbitration courts of Russian Federation: lexical and syntactic quality of texts: analytic note*. Saint Petersburg: IPP EUSPb, 2019. 20 p. (In Russian)

Слюсарь 2018 — Sl'usar N. A. *Experimental study of some basic concepts of theoretical morphology (on the material of the Russian language): regularity, syncretism, markedness. Author's abstract of the Doctor's Thesis.* Moscow, 2018. 48 p. (In Russian)

Солнышкина, Кисельников 2015 — Solnyshkina M. I., Kisel'nikov A. S. Text Complexity: Study Phases in Russian Linguistics. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiia.* 2015, (6 (38)): 86–99. (In Russian)

Azuelos-Atias, Ye 2017 — Azuelos-Atias S., Ye N. On drafting, interpreting, and translating legal texts across languages and cultures. *International Journal of Legal Discourse.* 2017, (2 (1)): 1–12.

Benjamin 2012 — Benjamin R. G. Reconstructing readability: recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*. 2012, (24 (1)): 63–88.

Blinova et al. 2020 — Blinova O., Belov S., Revazov M. Decisions of Russian Constitutional Court: Lexical Complexity Analysis in Shallow Diachrony. In: *CEUR Workshop Proceedings (Proceedings of the International Conference «Internet and Modern Society» IMS-2020)*. ITMO University, St. Petersburg, Russia. P. 61–74.

Blinova, Tarasov et al. 2020 — Blinova O. V., Tarasov N. A., Modina V. V., Blekanov I. S. Modeling Lemma Frequency Bands for Lexical Complexity Assessment of Russian Texts. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020" (Moscow, June 17–20, 2020)*. 2020, (19 (26)): 76–92.

Charrow, Charrow 1979 — Charrow R. P., Charrow V. R. Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions. *Columbia Law Review*. 1979, (79 (7)): 1306–1374.

Collins-Thompson 2014 — Collins-Thompson K. Computational assessment of text readability: a survey of current and future research. Recent Advances in Automatic Readability Assessment and Text Simplification. *Special issue of International Journal of Applied Linguistics.* 2014, (165 (2)): 97–135.

Ivanov et al. 2018 — Ivanov V. V., Solnyshkina M. I., Solovyev V. D. Efficiency of text readability features in Russian academic texts. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*. 2018, (17): 277–287.

Korobov 2015 — Korobov M. *Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts*. 320–332.

Mattila 2013 — Mattila H. E. S. *Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas*. Routledge, 2013. 504 p.

Owens, Wedeking 2011 — Owens R. J., Wedeking J. P. Justices and legal clarity: Analyzing the complexity of US Supreme Court opinions. *Law & Society Review*. 2011, (45 (4)): 1027–1061.

Solnyshkina et. al 2018 — Solnyshkina M., Ivanov V., Solovyev V. Readability Formula for Russian Texts: A Modified Version. In: *Proceedings of the 17th Mexican International Conference on Artificial Intelligence, MICAI 2018*. Guadalajara, Mexico, part II. P. 132–145.

Solovyev et al. 2018 — Solovyev V., Ivanov V., & Solnyshkina M. Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics. *Journal of Intelligent & Fuzzy Systems*. 2018, (34 (5)): 3049–3058.

Solovyev et al. 2020 — Solovyev V. D., Solnyshkina M. I., Andreeva M., Danilov A., Zamaletdinov R. Text Complexity and Abstractness: Tools for the Russian Language. IMS. In: *CEUR Workshop Proceedings (Proceedings of the International Conference «Internet and Modern Society» IMS-2020)*. ITMO University, St. Petersburg, Russia, 2020. P. 75–87.

Solovyev, Solnyshkina et al. 2019 — Solovyev V., Solnyshkina M., Ivanov V., Batyrshin I. Prediction of reading difficulty in Russian academic texts. *Journal of Intelligent & Fuzzy Systems*. 2019, (36 (5)): 4553–4563.

Straka, Straková 2019 — Straka M., Straková J. Universal Dependencies 2.5 Models for UDPipe (2019-12-06). In: *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2019.* URL: http://hdl.handle.net/11234/1-3131 (date of access: 01.02.2022).

Tiersma 1999 — Tiersma Peter M. *Legal Language*. Chicago, London: University of Chicago Press, 1999. 314 p.

Waltl, Matthes 2014 — Waltl B., Matthes F. *Towards Measures of Complexity: Applying Structural and Linguistic Metrics to German Laws*. JURIX. 2014: 153–162.

Wydick, Sloan 2019 — Wydick R. C., Sloan A. E. *Plain English for lawyers*. Sixth edition. Durham, North Carolina: Carolina Academic Press, LLC, 2019. 178 p.