Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2022»

Москва, 15–18 июня 2022 г.

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

Olga Blinova
St. Petersburg State University
St. Petersburg
7/9 Universitetskaya nab.,
o.blinova@spbu.ru

Nikita Tarasov
St. Petersburg State University
St. Petersburg
7/9 Universitetskaya nab.,
nkt.tarasov@yandex.ru

Abstract

The paper describes the metrics-based model for assessing complexity of Russian legal texts, implying 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, hypertext links”, “syntactic features”, “cohesion assessments”. The paper illustrates the reasons for choosing metrics, taking into account the experience of studies on linguistic complexity, stylometric studies, as well as experimental studies of legal texts perception. The authors present the results of testing the model in an experiment on the classification of texts by complexity level using metrics as parameters. These results are compared with the results of classification using USE (Universal Sentence Encoder) language model vectors as parameters. The authors come to the conclusion that the use of metrics makes it possible to assess text complexity more precisely than in an experiment using a language model.

Keywords: Russian legal texts; complexity assessment model; linguistic metrics; model testing
DOI: 10.28995/2075-7182-2022-21-1017-1028

Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности

Ольга Блинова
СПбГУ
Санкт-Петербург,
Университетская наб. 7/9
o.blinova@spbu.ru

Никита Тарасов
СПбГУ
Санкт-Петербург,
Университетская наб. 7/9
nkt.tarasov@yandex.ru

Аннотация

В статье описана основанная на метриках модель оценки сложности русских правовых текстов, подразумевающая использование 130 метрик следующих категорий: “базовые метрики”, “формулы читабельности”, “учёт слов разных частеречных классов”, “n-граммы частеречных тегов”, “частотность лемм”, “словообразовательные паттерны”, “отдельные граммемы”, “лексические и семантические признаки, неоднословные выражения, гипертекстовые связи”, “синтаксические признаки”, “оценки связности”. Иллюстрируются основания выбора метрик с учётом опыта исследований языковой сложности, стилеметрических исследований, а также экспериментальных работ в области восприятия правовых текстов. Приводятся результаты тестирования модели в эксперименте на классификацию текстов по уровню сложности с использованием полученных метрик в качестве параметров. Эти результаты сравниваются с результатами классификации с использованием в качестве параметров векторов языковой модели USE (Universal Sentence Encoder). Делается вывод, согласно которому использование метрик позволяет оценивать сложность текстов более точно, чем в эксперименте с использованием языковой модели.

Ключевые слова: Русские правовые тексты; модель оценки сложности; метрики сложности; тестирование модели

1 Введение

Юридические тексты (особенно – тексты законов) безотносительно правовой традиции и языка характеризуются как сложные, тёмные, запутанные и для неюриста непонятные, см. (Tiersma, 1999),

(Mattila, 2013), (Azuelos-Atias and Ye, 2017), (Bhatia, 1994) и мн. др. Настоящая статья посвящена описанию модели, разработанной для измерения **объективной сложности** [1] правовых текстов на русском языке. Наша модель, основанная на 130 метриках, разработана с учётом опыта исследований языковой сложности (в том числе – сложности юридических текстов), стилеметрических исследований, а также экспериментальных работ в области восприятия правовых текстов.

Задачи определения сложности текстов решаются достаточно давно. В том числе существует традиция применения методов оценки сложности к русским текстам, обзоры см., например, в (Reynolds, 2016), (Солнышкина, Кисельников, 2015). Наряду с понятием «сложность» в литературе используется понятие «читабельность». Читабельность понимается нами как оценка текста, полученная с применением параметров, которые в (Tuldava, 2017) названы латентными, в частности, формул читабельности и мер лексического разнообразия. Латентные параметры поддаются измерению, хотя и не поддаются непосредственному наблюдению в форме отдельных языковых сущностей, присутствующих в текстах. Соответственно, сложность мы понимаем как более комплексное явление, она оценивается и с обращением к скрытым параметрам, и с применением формально-статистических (поверхностных) параметров.

В Разделе 2 ниже мы иллюстрируем основания выбора метрик, в Разделе 3 кратко характеризуем типологию и набор метрик, в Разделе 4 излагаем методику и результаты тестирования модели. Более полное описание модели дано на сайте `https://www.plaindocument.org/`.

2 Мотивации выбора метрик

Сложность может пониматься как переменная, значение которой измеримо для любого (связного) текста на естественном языке. Модели оценки сложности развивались от простых (подразумевающих использование формул читабельности) к изощрённым (подразумевающим использование разнообразных метрик, обращающихся к лексике, морфологии, синтаксису, информации о частотности единиц текста и т. д.).

В нашей модели используются в том числе традиционные метрики сложности; прежде всего сказанное относится к категории базовых метрик и категории “формулы читабельности” (подробнее см. Раздел 3). Опыт, накопленный в работах по функциональной стилистике, стилеметрии и в психолингвистических исследованиях перцептивной сложности (трудности) при разработке нашей модели также учитывался.

Если учитывать, что **стилевая принадлежность текста коррелирует с его сложностью** (то есть в некотором общем случае деловые и научные тексты сложнее новостных, публицистических, разговорных), включение стилеспецифичных метрик получит обоснование.

Приведём некоторые примеры. В работах упоминается свойственный деловым текстам рост доли существительного и падение доли глагола в личной форме. Например, в (Голуб, 2001) указано, что “употребление глагольных форм сводится до минимума в официально-деловом стиле, который отличается наиболее ярко выраженным именным характером речи”, см. также (Кожина et al., 2011) и мн. др. Повышение доли существительных может объясняться по-разному.

Во-первых, принято упоминать о частотности в текстах официально-делового стиля (далее – ОДС) “глагольно-именных сочетаний» с “расщеплением сказуемого”, см. (Голуб, 2001) и мн. др., то есть о конструкциях с лёгкими глаголами типа ‘*оказывать содействие*’, ‘*производить замену*’. В нашей модели мы не только подсчитываем доли слов различных частей речи, но и учитываем вхождения конструкций с лёгкими глаголами.

Во-вторых, в литературе встречается суждение о частотности в текстах ОДС отглагольных номинализаций (безотносительно к их вхождениям в состав конструкций с лёгкими глаголами). Эта черта в составе модели учтена в словообразовательной метрике и (частично) в лексической метрике, учитывающей вхождения абстрактных слов.

[1] Вслед за (Dahl, 2004) различаются “сложность” (complexity) – объективная мера и “трудность” (difficulty) – мера субъективная. При сравнительной оценке текстов сложность может понимается как переменная, оказывающая влияние на восприятие текста читающим или слушающим (т. е. на трудность). Трудность в свою очередь – перцептивная характеристика. На трудность влияет не только объективная сложность текста, но и языковой (или, шире, – когнитивный) опыт воспринимающего текст субъекта.

В-третьих, увеличение доли существительного может объясняться через употребительность в ОДС неоднословных терминоподобных сочетаний типа '*товарищество собственников жилья*' (Дружкин, 2016). Эта черта учтена в лексической метрике, подсчитывающей вхождения юридических терминов (в том числе неоднословных).

В-четвёртых, доля существительных растёт за счёт неоднословных производных предлогов, компоненты которых размечаются как существительные, ср. '*в соответствии*', '*в связи*'. Эта черта также учтена в лексических метриках.

Представляется, что все четыре объяснения релевантны. Пример показывает, что учёт работ по стилистике позволяет анализировать сложность более подробным образом.

Практическая стилистика рекомендует не злоупотреблять пассивными конструкциями в некнижных стилях, см. (Голуб, 2001), а также (Wydick and Sloan, 2019) и мн. др. В работах о восприятии правовых текстов показано, что пассивные конструкции труднее активных, см., например, (Charrow and Charrow, 1979).

Соответственно, в модели среди метрик категории "отдельные граммемы" присутствует доля словоформ в творительном падеже (т.к. творительный падеж кодирует агенса в пассивных конструкциях). Кроме того, среди синтаксических метрик имеется доля вхождений пассивного подлежащего главной или зависимой клаузы. Наконец, учитывается доля личных форм глагола на *-ся*, а также (в составе частеречных метрик) – доля полных страдательных причастий и доля кратких страдательных причастий.

Важно заметить, что экспериментальные работы о сложности (точнее, перцептивной трудности) демонстрируют, что диагностическая сила некоторых традиционных метрик сложности для измерения актуальной понятности по данным эксперимента невысока. Например, в (Charrow and Charrow, 1979) показана невысокая предсказательная сила формул читабельности. Показано также, что длина предложения в стимуле практически не оказывала влияния на то, насколько успешно испытуемые справлялись с экспериментальным заданием по перефразированию, и что предложения одинаковой длины могут сильно различаться по фактической понятности.

Таким образом, сопоставление выводов квантитативных исследования сложности текстов и выводов экспериментальных исследований позволяет смотреть на предсказательную силу метрик сложности более трезво. В то же время эффективность метрик может быть проверена тестированием.

3 Набор метрик

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

1. базовые метрики;
2. формулы читабельности;
3. доли слов разных частеречных классов;
4. частотность лемм;
5. словообразование;
6. отдельные граммемы;
7. лексические и семантические признаки, неоднословные выражения, гипертекстовые связи;
8. синтаксические признаки;
9. оценки связности.

Полный список метрик и некоторые пользовательские словари доступны на сайте https://www.plaindocument.org/.

Модель предусматривает использование **28-ми базовых метрик**. Их можно разделить на базовые квантитативные и базовые лексические. Первые нацелены прежде всего на измерение длины слов и предложений (ср. 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, а также подсчёт долей гапаксов.

В модели используется **5 адаптированных для русского формул читабельности**: формула

Флеша-Кинкейда (Solnyshkina et al., 2018), SMOG, ARI, индекс Дейла-Чейл, индекс Колман-Лиау, см. (Бегтин, 2016).

22 метрики, учитывающие **доли вхождений слов разных частей речи**, разработаны с учётом различий между использованными в модели инструментами разметки. Для лемматизации, частеречной и синтаксической разметки использовался UDPipe (модель "ru-syntagrus") (Straka and Straková, 2019). Для второго слоя более подробной частеречной разметки и морфологической разметки использован pymorphy2 (Korobov, 2015). Под влиянием (Журавлев, 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). Кроме того, учитываются: отношение числа существительных к числу глаголов; доли сочинительных и подчинительных союзов; доли полных и кратких прилагательных; доли полных и кратких причастий; доля местоимений-существительных; доли предикативов, деепричастий, инфинитивов; доли числительных; доля частиц; доля однословных предлогов, а также доля форм компаратива.

Введены 13 метрик, обращающихся к представленности в текстах **n-грамм частеречных тегов**. Об эффективности метрик, учитывающих частеречную сочетаемость, см., например, (ping Tang and Cao, 2015). Отдельно 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 несколькими генитивными аргументами, которые 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» и следующий пример из Бюджетного кодекса РФ: *для обеспечения необходимой степени конфиденциальности рассмотрения отдельных разделов и подразделов расходов федерального бюджета и источников финансирования дефицита федерального бюджета Государственная Дума утверждает персональный состав рабочих групп <...>*.

Добавлена предложенная в (Антонова et al., 2011) "формула динамичности / статичности", призванная отделить тексты, в которых описывается множество событий ("динамические тексты") от текстов "статических". Эта метрика хорошо противопоставляет деловые тексты текстам других стилей (тексты официально-делового стиля более "статичны").

Использованы 9 метрик, учитывающих **вхождения лемм с разной общеязыковой частотностью**, принадлежащих 9-ти частотным диапазонам. Для подсчёта значений этой метрики на базе больших русских корпусов создан сводный частотный список лемм с индексами частотности Zipf value, см. (Blinova et al., 2020). Zipf value в этом списке принимает значения от 0 (наиболее низкочастотные леммы) до 8 (высокочастотные леммы). При оценке сложности учитываются доли вхождений в тексты лемм каждого из девяти частотных диапазонов.

Для диагностики сложности под влиянием (Дружкин, 2016) введена одна **словообразовательная метрика**. При подсчёте значений этой метрики модель обращается к уровню лемм, учитывая леммы вида **ция, *ние, *вие, *тие, *ист, *изм, *ура, *ище, *ство, *ость, *овка, *атор, *итор, *тель, *льный, *овать* (то есть подсчитывая вхождения некоторых отглагольных и отадъективных существительных, отглагольных прилагательных и производных глаголов). Заметим, что осложнённая когнитивная обработка производных слов по сравнению с непроизводными подтверждается в экспериментах на принятие лексического решения, см., например, (Нагель, 2017)

17 метрик категории "**отдельные граммемы**" заслуживают подробного обсуждения, мы вынужденно сократим изложение, приведя ограниченный ряд примеров. **Род существительных** учитывается, так как абстрактные существительные, употребительные в правовых текстах, часто среднего рода. **Граммема родительного падежа** хорошо диагностирует сложность, это известно из литературы вопроса, см., например, (Дружкин, 2016). **Творительный падеж** кодирует агенса в пассивных конструкциях. **Набор личных форм глагола** стилеспецифичен и жанровоспецифичен.

Согласно литературе вопроса, в ОДС частотны формы 3-го лица, формы 2-го лица практически не встречаются, а формы 1-го лица употребимы в ограниченном наборе жанров (Голуб, 2001).

11 метрик категории "**лексические и семантические признаки, неоднословные выражения, гипертекстовые связи**" также обращаются к описанным чертам текстов официально-делового стиля. Среди метрик категории: доля средств текстового дейксиса, обеспечивающих связность; доля графических сокращений; доля аббревиатур; доля леммы '*являться*'; доля юридических терминов; доля абстрактных лемм; доля лексических показателей деонтической возможности и необходимости; доля неоднословных предлогов; доля неоднословных оборотов в функции союза или союзного слова; доля конструкций с лёгкими глаголами, а также доля указаний на федеральные законы типа '*231-ФЗ*' (метрика призвана учитывать гипертекстовые связи).[2]

В 21-й **синтаксической метрике** учитываются:

- признаки, описывающие организацию отдельных синтаксических групп (именной группы – доля адъективных модификаторов имени; глагольной группы – доля наречных модификаторов предиката);
- признак, описывающий вхождения аппозитивных именных групп ("Appos");
- признаки, показывающих наличие сочинённых рядов (будь то сочинённые клаузы или однородные члены предложения; имеются в виду признак "Cc", описывающий союзные средства, а также признак "Conj", описывающий количество конъюнктов, в том числе вводимых бессоюзно);
- признаки, описывающие вхождения сентенциальных определений (причастий и причастных оборотов "Acl" и относительных клауз "Acl:relcl"), сентенциальных обстоятельств (деепричастий и зависимых клауз с личными формами глагола, "Advcl"), различных сентенциальных дополнений ("Ccomp", "Xcomp"), а также так называемых конструкций с сентенциальным субъектом ("Csubj", "Csubj:pass"); отдельно учитываются единицы, способные вводить зависимые клаузы ("Mark");
- признак, описывающий вхождения клауз со связочными элементами ("Cop");
- признаки, с разных точек зрения описывающие вхождения пассивных конструкций ("Aux:pass", "Nsubj:pass", "Csubj:pass").

Наконец, 2 **метрики связности** оценивают количество повторов существительных в соседних предложениях и количество повторов граммем времени и вида у глаголов в личной форме (в соседних предложениях).

4 Тестирование модели

Для определения качества выбранных 130 метрик, их способности предсказывать сложность текстов произведены такие тесты и сравнения:

- классификация с использованием полученных метрик в качестве параметров;
- классификация с использованием в качестве параметров векторов языковой модели.

4.1 Тестирование на текстовом наборе "plainrussian"

Тесты проводились на стандартном текстовом наборе "plainrussian" И. Бегтина, включающем тексты, распределённые на группы по уровню образования (с 3-го класса начальной школы до 6 курса вуза) (Бегтин, 2016). Из-за ограниченного размера тестового набора (68 текстов) для тестирования данные были разбиты на 3 класса: "простые тексты" – до 6-го класса, "средние по сложности тексты» – с 6 по 11 классы, "сложные тексты" – тексты уровня высшего образования. Итоговое число

[2](Azuelos-Atias and Ye, 2017) указывают, что неудобопонятность правовых текстов для неспециалистов обуславливает структурная сложность (прежде всего — синтаксическая), понятийная сложность (употребление специальной терминологии), а также "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". Таким образом, можно говорить о своеобразной "интертекстуальной сложности", которая может оцениваться через подсчёт количества ссылок на "внешние" тексты, необходимые для интерпретации содержания читаемого текста. Наша метрика, оценивающая гипертекстовые связи, введена именно из этих соображений.

USE кодировки			
Тип текста	**Точность**	**Полнота**	**F-мера**
Простой текст	0,506	0,583	0,524
Текст средней сложности	0,667	0,333	0,419
Сложный текст	0,634	0,736	0,679
Кодировки метриками			
Тип текста	**Точность**	**Полнота**	**F-мера**
Простой текст	0,778	0,806	0,775
Текст средней сложности	0,567	0,733	0,622
Сложный текст	0,849	0,778	0,811

Таблица 1: Оценки классификации в эксперименте с “plainrussian”.

документов для каждой группы: “простые” – 14, “средние” – 32, “сложные” – 22. В качестве тестовой модели классификации использован XGBoost (Chen and Guestrin, 2016).

4.2 Классификация с использованием в качестве параметров векторов языковой модели

Сравнение производилось с языковой моделью USE (Universal Sentence Encoder) (Cer et al., 2018) с использованием современной нейросетевой архитектуры “Transformer”. Оно позволило получить представление об эффективности выбранных метрик в задаче классификации по сложности. Таким способом проверено качество кодирования сложности текстов в описанном подходе по сравнению с подходом, кодирующим тексты на основе выбранных 130 метрик, отражающих знания о естественном языке.[3]

Модель тестировалась с предварительным разбиением на тестовую и тренировочную выборки с последующим подбором гиперпараметров с помощью библиотеки “Hyperopt” (Bergstra et al., 2013). Для подбора параметров было обучено 1000 моделей с различными параметрами.

Цитируемые выше показатели качества (см. Таблицу 1) приводятся для оптимизированной модели с использованием кросс-валидации (Refaeilzadeh et al., 2016) с разбиением данных на 10 групп. Этот подход даёт возможность показать результаты более объективно, учесть генерализацию модели для ранее не использованных данных, что особенно важно в случае работы с небольшими наборами данных.

Таким образом, метрики позволяют получить более точные оценки сложности текстов. Наиболее успешно выделяются “сложные тексты”, несколько менее успешно – “простые тексты”, наименее успешно – “тексты средней сложности”.

4.3 Тестирование на текстовом наборе учебников обществознания

Вторая итерация тестов проводилась проводились на наборе учебников обществознания, распределённые на группы по классам общеобразовательной школы (5 – 11 классы) (Solovyev et al., 2018). Данные также были распределены на 3 категории: “более простые тексты»” – 5, 6, 7 классы, “средние по сложности тексты” – 8, 9 классы, “более сложные тексты” – 10, 11 классы. Итоговое число документов для каждой группы: “более простые” – 5, “средние” – 4, “более сложные” – 5, размер датасета – 716 тыс. слов, средняя длина документа – приблизительно 1200 строк (по предложению на строку).

Все документы были случайным образом разбиты на фрагменты длиной в 100 строк. Затем данные были размечены с использованием UDPipe и pymorphy2, для каждого фрагмента вычислены значения 130 метрик. После этого была выполнена классификация. В качестве тестовой модели классификации использован XGBoost (Chen and Guestrin, 2016).

[3]Применение предобученной нейросетевой модели оправдано малым объемом данных. В случае использования классических подходов к кодированию существует риск не получить точные представления текстов, в частности, потому, что многие слова могут встретиться лишь в нескольких документах, существенно ухудшая качество частотных моделей (таких, как tf-idf). Использование предобученных текстовых моделей позволяет обойти это ограничение, т. к. в них языковая модель обучается на больших объемах данных.

Кодировки метриками			
Тип текста	**Точность**	**Полнота**	**F-мера**
Простой текст	0,929	0,867	0,897
Текст средней сложности	0,793	0,920	0,852
Сложный текст	0,971	0,895	0,932

Таблица 2: Оценки классификации в эксперименте с учебниками.

Итоговые показатели качества для кодирования с использованием метрик приведены в Таблице 2.

В описанных экспериментах получены данные об эффективности работы 130 метрик в задаче классификации по сложности. Тестирование проводилось на наборах данных, существенно отличающихся от наших. Между тем, некоторые метрики были целенаправленно разработаны для применения к текстам ОДС. В текстах других стилей по крайней мере некоторые учитываемые нами признаки могут описывать редкие или сверхредкие явления.

4.4 Эффективность отдельных метрик

Эксперимент с “plainrussian” показал, что в задаче классификации эффективны 72 метрики. В эксперименте с учебниками обществознания выяснилось, что для классификации важна прежде всего формула Флеша-Кинкейда, коэффициенты (константы) которой вычислялись как раз на датасете с учебниками обществознания его создателями (Solovyev et al., 2018), а также 94 других признака. В наборах работающих на классификацию по сложности признаков совпадает 56, см. Приложение 1, где дан список признаков, расположенных по убыванию их суммарной значимости.

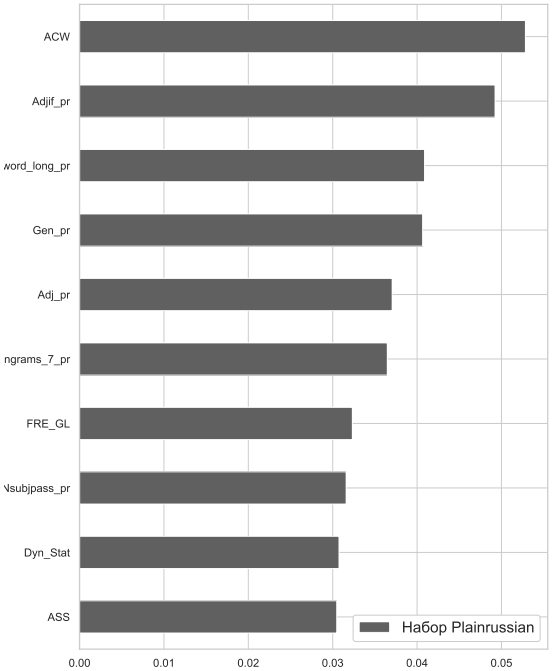

Рис. 1: Топ-10 метрик, “plainrussian”.

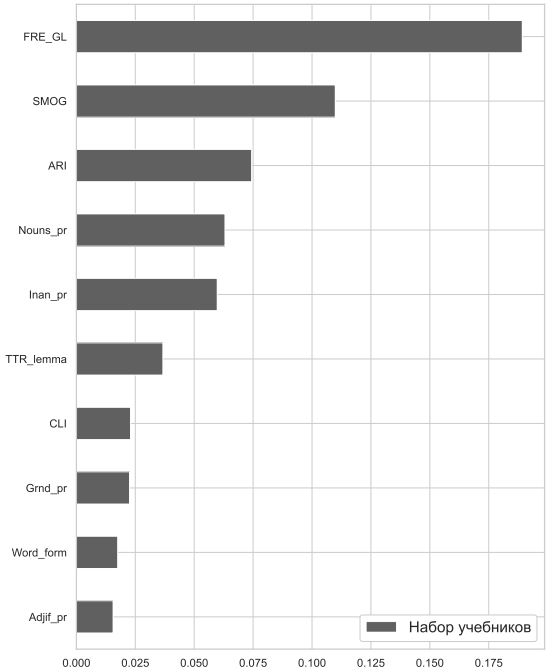

Рис. 2: Топ-10 метрик, учебники.

[4] В число десяти наиболее эффективных метрик в эксперименте с “plainrussian” вошли: средняя длина словоформы в буквах, доля полных прилагательных, доля слов длиной 4 и более слога, доля словоформ в родительном падеже, доля прилагательных, доля биграмм тегов существительного и существительного в род. п., формула Флеша-Кинкейда, доля вхождений пассивного подлежащего главной или зависимой клаузы, формула динамичности / статичности и средняя длина предложения в слогах, см. Рис. 1.

На классификацию текстов учебников (см. Рис. 2) лучше других метрик сработали: формулы читабельности (FRE_GL, SMOG, ARI),[5] а также индекс именной лексики, доля неодушевлённых существительных, индекс Колман-Лиау, доля лемм с «хвостами» типа **ция, *ние, *вие, *тие, *ист* (см. о них Раздел 3 выше), доля полных прилагательных, доля кратких прилагательных и доля адъективных модификаторов имени.

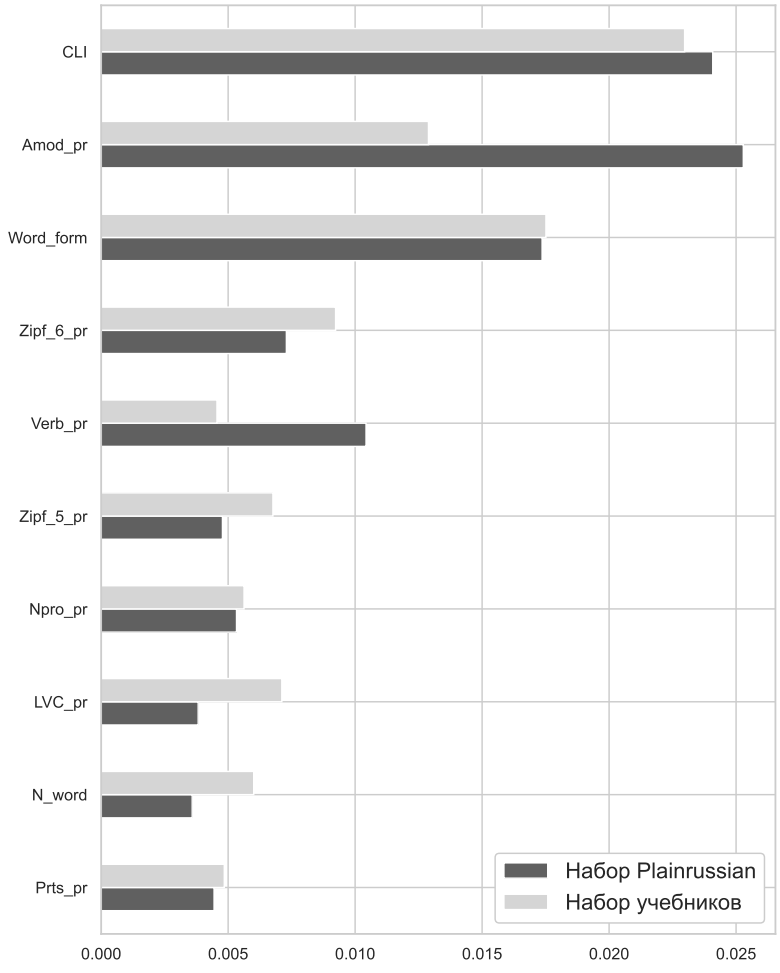

Рис. 3: Топ-10 метрик по суммарной значимости.

[4]Значимости параметров для итоговой модели классификации были получены с использованием стандартной функции “Importance” библиотеки XGBoost. Более высокие значения говорят о большем влиянии метрики на выбор класса.

[5]Высокая эффективность FRE_GL неудивительна: текстовой набор, на которых мы проверяли эффективность метрик, –- это и есть набор, на котором формула была адаптирована.

На Рис. 3 представлены метрики, эффективные в обоих экспериментах. Они ранжированы по суммарной значимости и отобраны так: вес каждого из элементов (т. е. метрики для определённого набора данных) не превышаает 70% от общей суммы. Среди них (в порядке убывания значимости): индекс Колман-Лиау, доля адъективных модификаторов имени, доля лемм с «хвостами», включающими определённые словообразовательные аффиксы, доля среднечастотных лемм (Zipf value = 6), индекс глагольности, доля среднечастотных лемм (Zipf value = 5), доля местоимений-существительных, доля конструкций с лёгкими глаголами, количество словоформ и доля кратких причастий.

5 Перспективы

В настоящей статье охарактеризована модель оценки сложности, в которой учитывалось 130 параметров, в том числе – стилеспецифичных (т.е. целенаправленно выделенных для русских текстов ОДС). Продолжением работы станет превращение модели, основанной на метриках, в гибридную. Использование метрик в совокупности с эффективным кодированием языка позволит оценивать сложность как по языковым параметрам, так и по неявным признакам.

При тестировании модели мы столкнулись с нехваткой доступных русскоязычных текстовых наборов с оценками сложности (читабельности), содержащих изучаемые нами тексты. Мы использовали набор “plainrussian”, содержащий в общей сложности 68 текстов, а также существенно более обширный датасет из 14-ти учебников (Solovyev et al., 2019). Таким образом, тестирование проводилось на наборах данных, существенно отличающихся от наших.

Нашей ближайшей задачей станет создание текстового набора ОДС с оценками сложности, полученными не с помощью метрик.[6] Наличие такого набора позволит адаптировать формулы читабельности для русских юридических текстов. Формулы дадут возможность измерять читабельность правовых текстов, не подвергшихся разметке.

Представленность в текстах различных перечисляемых выше признаков может иметь и положительную, и отрицательную корреляцию с целевой сложностью. Оценить корреляции мы сможем, имея упомянутый текстовой набор ОДС. Наконец, нам предстоит определить и использовать стратегию подсчёта агрегированного индекса сложности на основе значений всех метрик.

Благодарности

Авторы хотели бы поблагодарить Валерия Дмитриевича Соловьёва за доступ к датасету школьных учебников. Исследование выполнено при поддержке гранта РНФ № 19-18-00525.

References

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

James Bergstra, Dan Yamins, David D. Cox, et al. 2013. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. // *Proceedings of the 12th Python in science conference*, volume 13, P 20. Citeseer.

Vijay K. Bhatia. 1994. Cognitive structuring in legislative provisions. // John Gibbons, *Language and the Law*, P 136–155. Longman.

Olga Blinova, Nikita Tarasov, Valerija Modina, and Ivan Blekanov. 2020. Modeling lemma frequency bands for lexical complexity assessment of russian texts. // *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2020)*, volume 19(26), P 76–92.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

[6] Мы планируем предъявлять читающим фрагменты текстов равной длины и получать эксплицитные экспертные оценки сложности, попутно измеряя время чтения.

Robert P. Charrow and Veda R. Charrow. 1979. Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review*, 79(7):1306–1374.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. // *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, P 785–794.

Östen Dahl. 2004. *The growth and maintenance of linguistic complexity*. John Benjamins Publishing, Amsterdam.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. // Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry Ignatov, and Valeri G. Labunets, *Analysis of Images, Social Networks and Texts*, P 320–332, Cham. Springer International Publishing.

Heikki E. S. Mattila. 2013. *Comparative legal linguistics: language of law, Latin and modern lingua francas*. Ashgate Publishing, Ltd., Farnham, Surrey, 2 edition.

Xiao ping Tang and Jing Cao. 2015. Automatic genre classification via n-grams of part-of-speech tags. *Procedia - Social and Behavioral Sciences*, 198:474–478.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2016. Cross-validation. *Encyclopedia of database systems*, P 1–7.

Robert Reynolds. 2016. Insights from russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. // *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, P 289–300, San Diego, California.

Marina Solnyshkina, Vladimir Ivanov, and Valery Solovyev. 2018. Readability formula for russian texts: A modified version. // *Proceedings of the 17th Mexican International Conference on Artificial Intelligence, MICAI 2018*, P 132—145, Guadalajara, Mexico.

Valery Solovyev, Vladimir Ivanov, and Marina Solnyshkina. 2018. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent Fuzzy Systems*, 34:3049–3058.

Valery Solovyev, Marina Solnyshkina, Vladimir Ivanov, and Ildar Batyrshin. 2019. Prediction of reading difficulty in Russian academic texts. *Journal of Intelligent Fuzzy Systems*, 36:4553–4563.

Milan Straka and Jana Straková. 2019. Universal dependencies 2.5 models for UDPipe (2019-12-06). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Peter M. Tiersma. 1999. *Legal Language*. The University of Chicago Press, Chicago, London.

Juhan Tuldava. 2017. The development of statistical stylistics (a survey). *Journal of Quantitative Linguistics*, 11(1–2):141–151.

Richard C. Wydick and Amy E. Sloan. 2019. *Plain English for lawyers*. Carolina Academic Press, LLC, Durham, North Carolina, 6 edition.

А. Ю. Антонова, Э. С. Клышинский, Е. В. Ягунова. 2011. Определение стилевых и жанровых характеристик коллекций текстов на основе частеречной сочетаемости. // *Труды международной конференции «Корпусная лингвистика–2011»*, P 80—85, Санкт-Петербург. СПбГУ.

И. Бегтин. 2016. Plainrussian.ru. `https://github.com/ivbeg/readability.io`.

И. Б. Голуб. 2001. *Стилистика русского языка*. Рольф, Москва.

К. Ю. Дружкин. 2016. *Метрики удобочитаемости для русского языка*. выпускная кфалификационная работа магистра, НИУ ВШЭ, Москва.

А. Ф. Журавлев. 1988. Опыт квантитативно-типологического исследования разновидностей устной речи. // *Разновидности городской устной речи*, P 84—150. Наука.

М. Н. Кожина, Л. Р. Дускаева, В. А. Салимовский. 2011. *Стилистика русского языка*. Флинта, Наука, Москва.

О. В. Нагель. 2017. *Словообразовательные механизмы в процессах восприятия, идентификации и использования языка*. автореф. дисс. … докт. филол. наук, Томский государственный университет, Томск.

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

Приложение 1. Список совпадающих признаков и их значимость

метрика	пояснение	учебники	plainrussian
FRE_GL	адаптированная формула Флеша-Кинкейда	0,1892	0,0323
SMOG	адаптированная формула SMOG	0,1099	0,0187
ARI	адаптированная формула подсчёта автоматизированного индекса читабельности	0,0744	0,0221
Nouns_pr	индекс именной лексики	0,0631	0,0045
Inan_pr	доля неодушевлённых существительных	0,0598	0,0064
Adjif_pr	доля полных прилагательных	0,0156	0,0492
ACW	средняя длина словоформы в буквах	0,0033	0,0528
Gen_pr	доля словоформ в родительном падеже	0,0124	0,0407
CLI	индекс Колман-Лиау	0,023	0,0241
word_long_pr	доля длинных слов (4 и более слога)	0,0009	0,0409
Adj_pr	индекс адъективности	0,0035	0,037
Amod_pr	доля адъективных модификаторов имени	0,0129	0,0253
Nsubjpass_pr	доля вхождений пассивного подлежащего главной или зависимой клаузы	0,0064	0,0316
ASS	средняя длина предложения в слогах	0,0049	0,0305
Word_form	доля лемм с «хвостами», включающими определённые словообразовательные аффиксы (или их фрагменты)	0,0175	0,0174
Dyn_Stat	формула динамичности / статичности	0,0036	0,0307
Prtf_pr	доля полных причастий	0,0045	0,0297
Abstr_pr	доля абстрактных лемм	0,0073	0,0234
Pos_ngrams_1_pr	доля биграмм тегов глагола в личной форме и существительного	0,0004	0,0281
DCI	индекс Дейла-Чейл	0,0053	0,0196
ASW	средняя длина словоформы в слогах	0,0012	0,0227
Prep_mw_pr	доля неоднословных предлогов	0,0062	0,0146
Aclrelcl_pr	доля относительных клауз	0,0034	0,017
Pos_ngrams_11_pr	доля биграмм тегов существительного и полного причастия	0,0006	0,0168
Adjs_pr	доля кратких прилагательных	0,0154	0,0018
lemma_long_pr	доля длинных лемм	0,0039	0,0132
Abbr_pr	доля аббревиатур	0,0121	0,0044
Zipf_6_pr	доля среднечастотных лемм (Zipf value = 6)	0,0092	0,0073

Acl_pr	доля клаузальных модификаторов имени	0,0034	0,0126
Verb_pr	индекс глагольности	0,0046	0,0104
TTR_word	простой TTR (словоформы)	0,0109	0,0039
Nummod_pr	доля числовых модификаторов существительного	0,003	0,0118
N	количество числовых символов	0,0019	0,0107
Cop_pr	доля клауз с элементами, трактуемыми как связочные	0,0084	0,0034
Zipf_5_pr	доля среднечастотных лемм (Zipf value = 5)	0,0068	0,0048
LVC_pr	доля конструкций с лёгкими глаголами	0,0071	0,0038
Npro_pr	доля местоимений-существительных	0,0056	0,0053
V_lemma	количество типов (леммы)	0,0014	0,0093
N_word	количество токенов	0,006	0,0036
Prts_pr	доля кратких причастий	0,0048	0,0044
sent	количество предложений	0,0042	0,0044
Cohes_2	количество повторов граммем времени и вида у глаголов в личной форме (в соседних предложениях)	0,0043	0,0042
Prcl_pr	доля частиц	0,0018	0,0067
Cc_pr	доля союзов, связанных с конъюнктами синтаксическим отношением ”cc”(координации)	0,0047	0,0032
Prep_pr	доля однословных предлогов	0,0052	0,0024
Nsubj_pr	доля вхождений активного подлежащего главной или зависимой клаузы	0,0001	0,0068
NVR	Noun-Verb ratio	0,0006	0,0062
Yavl_pr	доля леммы «являться»	0,0004	0,0054
Advcl_pr	доля обстоятельственных клауз	0,0018	0,0039
Ablt_pr	доля словоформ в творительном падеже	0,0001	0,0043
C	количество знаков	0,0002	0,0039
Zipf_4_pr	доля среднечастотных лемм (Zipf value = 4)	0,0025	0,0012
Impf_pr	доля глаголов несовершенного вида	0,003	0,0006
Pssv_prtf_pr	доля полных страдательных причастий	0,0019	0,0011
Zipf_3_pr	доля низкочастотных лемм (Zipf value = 3)	0,0018	0,0012
Advmod_pr	доля наречных модификаторов (наречий или наречных групп)	0,0008	0,0007