Правительство Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет»
(СПбГУ)

УДК 81"42; 801.7
Рег. № НИОКТР
АААА-А19-119050790087-9
Инв. № 72674776

УТВЕРЖДАЮ
Начальник Управления
научных исследований СПбГУ

_____ Е.В. Лебедева
«» ____2022 г.

ОТЧЁТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

Понятность официального русского языка: юридическая и лингвистическая проблематика

(заключительный)

по гранту РНФ
№ 19-18-00525 от 01.04.2019

Руководитель НИР,
доцент,

кандидат филологических наук О.В. Блинова

Санкт-Петербург
2022

Реферат

Проект «Понятность официального русского языка: юридическая и лингвистическая проблематика» нацелен на изучение языковой сложности русских официально-деловых (правовых, юридических) текстов.

В рамках исследования последовательно различаются «**сложность**» (complexity) – абстрактная объективная мера и «**трудность**» (difficulty) – мера относительная. При сравнительной оценке текстов, стилей, регистров сложность может пониматься как текстовая переменная, оказывающая влияние на восприятие текста читающим или слушающим (т. е. на трудность). Трудность в свою очередь – перцептивная характеристика текста, стиля или регистра. На трудность влияет не только объективная сложность текста, но и языковой (или, шире, – когнитивный) опыт воспринимающего текст субъекта.

Кроме того, в исследовании задействована «**понятность**» (clarity); этим понятием оперируют прежде всего юристы, описывая, насколько правовой текст доступен для так называемого «простого гражданина» (неюриста), и каковы юридические последствия непонимания. Понятность (как и трудность) можно считать функцией от сложности текста и опыта адресата.

В рамках проекта разработана автоматическая **модель оценки сложности**, в которой используется 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 (vague) контексты.

В рамках изучения **трудности** официальных текстов проведён масштабный социолингвистический опрос и серия интервью.

Изучение **понятности** велось в ходе мониторинга сайтов государственных органов. Мониторинг официальных сайтов позволил оценить соответствие содержания официальных сайтов требованиям доступности информации о деятельности государственных органов и органов местного самоуправления, а также рассмотреть информационное содержание официальных сайтов с точки зрения его понятности адресату – пользователю сайта.

Кроме того, в рамках проекта сформирована «Концепция использования государственного языка в деятельности государственных и муниципальных органов и организаций» (см. Приложение к настоящему отчёту).

Таким образом, исследование велось в рамках трёх основных направлений («**дескриптивного**», или корпусного, направленного на изучение сложности; «**перцептивного**», направленного на изучение трудности и обработку данных массового социолингвистического опроса; «**мониторингового**», направленного на изучение понятности и описание контента сайтов государственных органов и организаций). Объектом исследования являлся русский правовой текст; он исследовался методами квантитативной лингвистики, через обращение к носителям и с применением специальных юридических знаний.

Содержание

Введение

Сложность правовых текстов исследуется с применение квантитативных (корпусных) методов. В этой области обычно не предлагается каких-то пионерских подходов, специально разработанных для юридических текстов; авторы работ следуют классическим методикам, выбирая вполне традиционные метрики текстов и параметры оценки сложности (среди них: формулы читабельности, длина предложений и слов текста, общеязыковая частотность слов текста и т. д.). Дополнением классических подходов является внимание к явлениям **неопределённости (vagueness)** и к **гипертекстовым связям** внутри изучаемых корпусов.

Так, в [Waltl, Matthes, 2014] при оценке сложности законов используются в том числе отношения, в которые вступает конкретное положение правового текста, содержащее ссылку на другое положение этого же или другого правового текста (назову это «внутризаконодательным дейксисом»).

В целом авторы [Там же] использовали следующие метрики оценки сложности: количество абзацев, количество предложений, количество слов, «структурная глубина» (параметр не имеет отношения к языковым структурам, но описывает организацию корпуса немецких законов и текстовые ссылки разной «глубины» на тот же законодательный акт/другой акт, это оценка «внутризаконодательного дейксиса»), количество внутренних и внешних ссылок; разнообразие словаря, значение стандартной метрики читабельности. [Waltl, Matthes, 2014] включили в оценку сложности «неопределённость», вычисляемую так: оценивалось количество вхождений в тексты немецких законов неопределённых (vague) прилагательных типа «*адекватный*», «*разумный*»; о неопределённости в русских правовых текстах см. [Блинова, Белов, 2020].

Таким образом, при оценке понятности юридических текстов исследователи зачастую анализируют явления языковой **неопределённости** (vagueness) в текстах в дополнение к их сложности в традиционном понимании (длинные слова и предложения, количество подчинённых клауз и проч.), см. [Owens, Wedeking, 2011]. Скажем, в [Там же] авторы, разрабатывая свою аналитическую модель (Linguistic Inquiry and Word Count), учитывали **10 параметров «когнитивной сложности»**, среди которых:

1. “causation” (измерение количества выраженных в тексте причинно-следственных связей через поиск слов типа ‘*because*’, ‘*effect*’, ‘*hence*’);
2. “insight” (измерение трудноинтерпретируемого показателя, смысл которого авторы формулируют так: “the degree to which individuals differ in how much each is able to discern a more in-depth understanding of a subject or its underlying nature”; оценивается через нахождение в текстах слов типа ‘*think*’, ‘*know*’, and ‘*consider*’; вероятно, имеются в виду прежде всего эпистемические глаголы в роли хеджей, см. о них [Блинова, Белов, 2020]);
3. “discrepancy” (измерение степени выявления пишущим субъектом несоответствий, различий между описываемыми в тексте ситуациям; оценивается через нахождение в тексте слов типа ‘*should*’, ‘*would*’, ‘*could*’);
4. “inhibit” (измерение количества запретов, выраженных пишущим субъектом, принимающим решение; оценивается через нахождение в тексте слов типа ‘*block*’, ‘*stop*’);

5. “tentativeness” (измерение степени неуверенности пишущего субъекта; оценивается через нахождение в тексте слов типа ‘*maybe*’, ‘*fairly*’, ‘*perhaps*’, то есть, собственно, хеджей);

6. “certainty” (измерение степени уверенности пишущего субъекта; оценивается через нахождение в тексте слов типа ‘*always*’, ‘*absolutely*’, ‘*clearly*’, то есть хеджей);

7. “inclusiveness” (измерение выраженности связей между сущностями в тексте; оценивается через нахождение в тексте слов типа ‘*with*’, ‘*and*’, то есть, по-видимому, сочинённых и комитативных конструкций);

8. “exclusiveness” (измерение количества выражений принадлежности/непринадлежности описываемой в тексте сущности в какой-либо категории; оценивается через нахождение в тексте слов типа ‘*but*’, ‘*except*’);

9. “negations” (измерение трудноинтерпретируемого показателя, “measure to what extent an individual acknowledges the absence or opposite of something that is positive or affirmative”; оценивается через нахождение в тексте слов типа ‘*no*’, ‘*never*’);

10. “six-letter” (доля слов текста, содержащих шесть и более букв; выделение порога в шесть букв авторами не поясняется).

Несмотря на некоторую невнятность статьи [Owens, Wedeking, 2011] достаточно ясно, что из 10 параметров лишь один оценивает сложность в традиционном понимании; остальные направлены на выявление неопределённости (vagueness) и оценку некоторых других текстовых показателей «когнитивной сложности».

Оценка сложности находит в праве интересные применения, например, [Zubrod et al., 2020] рассматривали использование языка на судебных процессах и оценили связь между оценками языковой сложности и результатами судебных разбирательств (обнаружив, что более высокие уровни сложности коррелируют со значительным увеличением выигрышных исходов, но только для стороны обвинения).

Об исследованиях, направленных на **оценку сложности русских правовых текстов**, стоит рассказать подробнее. Русские тексты привлекали внимание исследователей, которые, во-первых, сконцентрировались в основном на оценке сложности текстов **законов**, во-вторых, использовали для оценки сложности любо только **формулы читабельности**, либо другие достаточно **простые и немногочисленные метрики**.

Одной из первых работ в этой области является статья [Костенко, 2005], которая содержит скорее общие рассуждения, чем описание конкретных подходов. Автор предлагает оценивать «лингвистическую корректность» нормативно-правовых актов. Первый этап заключается «в детальном анализе каждой из словоформ, входящих в состав предложения, и разбиения предложения на смысловые единицы», что предполагает изучение юридических терминов и понятий и определение их «семантического окружения». Второй этап подразумевает проверку «соблюдения свойства проективности фраз естественного языка, которое отражает правильность <sic – О.Б.> синтаксической структуры соответствующего предложения текста». Вопрос о том, какое отношение может иметь имеет оценка проективности к анализу юридических текстов, ставить стоит; однако теоретические ожидания не позволяют ожидать, что в правовых текстах с какой-то периодичностью встречаются непроективные структуры, хотя синтаксически предложения могут быть устроены весьма сложно. Например, в [Голуб, 2001] о линейном порядке слов

читаем: «В научном и официально-деловом стилях, как правило, порядок слов не используется в экспрессивной функции и потому инверсия не может быть оправдана».

Третий этап гипотетического анализа включает оценку текстов по параметру «сложности или громоздкости предложений текстов нормативно-правовых актов» [Костенко, 2005]. Здесь высказывается внятная мысль о том, что чем проще текст, тем больше он содержит точек и тем меньше запятых. Кроме того, автор упоминает о гипотезе глубины [Ингве, 1965], о подходах к оценке синтаксической сложности [Мартыненко, 1971] и заключает, что «лингвистическая сущность понятия громоздкости предложений объясняется достаточно простым и известным явлением, которое приводит к осложнению порядка слов. Данное явление может быть вызвано: а) дистантным расположением пары «управляемое-управляющее»; б) удалением зависимых слов от своего управляющего; в) гнездованием, которое приводит к специфичному разрыву синтаксически связанных слов <…>. Таким образом, громоздкость предложений возрастает с увеличением протяженности или длины дуги (явление дистантности), а также с увеличением степени обрамления дуг (явление гнездования)». К сожалению, других работ М. А. Костенко найти не удалось, остаётся неясным, удалось ли автору применить предложенных подходы оценки нормативно-правовых текстов на практике.

В статье [Дмитриева, 2017] корпус текстов решений Конституционного суда РФ исследован с применением простой метрики оценки читабельности – формулы Флеша-Кинкейда, адаптированной для русского языка И.В. Оборневой [Оборнева, 2005], в которой учитываются ASL — средняя длина предложения в словах (рассчитываемая как количество слов документа, делённое на количество предложений документа), ASW – средняя длина слова в слогах (рассчитываемая как общее количество слогов в тексте документа, делённое на общее количество слов документа). Относительно [Дмитриева, 2017] стоит заметить следующее. Коэффициенты формулы Оборневой были получены в результате вычисления статистических характеристик около 100 произведений известных англоязычных литературных классиков (и переводов этих произведений на русский язык). Таким образом, формула не вполне универсальна, а применима прежде всего для анализа сложности текстов (переводной) художественной литературы, об указанной проблеме см. [Solnyshkina et. al 2018].

Изучением сложности занимаются и сотрудники Института проблем правоприменения при Европейском университете в Санкт-Петербурге Денис Савельев и Руслан Кучаков, см. [Кучаков, Савельев, 2018], [Савельев, Кучаков, 2019]. В цитированных исследованиях для оценки сложности авторы использовали одну **метрику лексического разнообразия** (TTR, значение которой не независимо от длины текста, следовательно, результаты применения метрики могут подвергаться сомнению) и одну синтаксическую метрику (maximum Dependency Length, maxDepLen), расстояние между главным и зависимым по синтаксическому дереву зависимости, вычисляемое так: «для каждого конкретного текста взято **одно значение, которое является максимальным для всех предложений текста**» [Там же]). В [Савельев, 2020] метрики не перечислены в виде эксплицитного списка, но, похоже, к TTR и maxDepLen были добавлены длина предложения в словах и анализ количества вхождений токенов, «которые представляют собой самостоятельные части речи: существительные, прилагательные, глаголы и наречия».

В новой работе [Кнутов и др., 2020] использовано большее количество метрик (девять), они вполне традиционны:

1. «доля глаголов в страдательном залоге»,
2. «доля глаголов от общего количества слов в тексте»,
3. «среднее количество слов в субстантивных именных словосочетаниях»
4. «среднее количество причастных оборотов, расположенных в предложениях после определяемого слова, на одно предложение»
5. «среднее количество деепричастных оборотов на одно предложение»
6. «среднее количество слов в предложениях»,
7. «среднее расстояние между зависимыми словами в предложении»,
8. «среднее количество грамматических основ (предикативных основ, предикативных ядер) предложения (подлежащее, сказуемое или одно из них) в одном предложении»,
9. «среднее количество слов в абзаце».

К сожалению, авторы цитируемой работы не эксплицируют основания выбора параметров; между тем, читателю они понятны не всегда. Скажем, неясно, почему именно расположение причастных оборотов «после определяемого слова» — это показатель сложности. Не совсем понятно, что имеется в виду под «долей глаголов в страдательном залоге», вероятно, только пассивные причастия (так как финитным формам глагола граммемы залога на слое морфологической разметки не приписываются).

[Кнутов и др., 2020, 44-45] определили, насколько значение параметров влияет на усложнение текста, и получили следующие показатели, цитирую: «Если предположить, что использование всех негативных практик усложняет восприятие текста на 100%, то доли каждой из таких практик будут следующие: 1) малое количество глаголов – 18%; 2) большое количество слов в субстантивных именных словосочетаниях — 16%; 3) большое количество слов в абзацах – 15%; 4) большое расстояние между зависимыми словами – 15%; 5) большое количество причастных оборотов, расположенных после определяемого слова – 14%; 6) большое количество слов в предложении – 11%; 7) большое количество грамматических основ (предикативных основ, предикативных ядер) предложения (подлежащее, сказуемое или одно из них) – 11%».

Таким образом, в прикладных исследованиях сложности текстов накоплен важный опыт, в том числе есть опыт оценки сложности русских правовых текстов. В описываемом проекте он учтён.

Общедоступных размеченных корпусов русских официально-деловых текстов с онлайн-поиском пока не существует, их создание – дело будущего [Крылов, Фролова, 2017]. При этом электронные коллекции документов относительно многочисленны. Это, в частности, собрания нормативных документов на сайте РОМИП (документы Законодательства РФ, Москвы и Санкт-Петербурга, см. [Российский семинар по оценке методов информационного поиска]), свод законов Российской империи Российской национальной библиотеки [Полное собрание законов Российской империи], библиотека нормативно-правовых актов СССР [Библиотека нормативно-правовых актов СССР] и др.

Вообще говоря, в цифровом мире существует огромное количество русскоязычных юридических текстов, они доступны, например, через поиск в правовых информационных системах «Консультант Плюс», «Гарант», «Континент», «Законодательство» и некоторых

других. Есть и коллекции размеченных данных, например, “RusLawOD: Russian Law Open Data” Д. Савельева.

Официально-деловые тексты вошли в состав некоторых корпусов русского языка. Так, в «Машинном фонде русского языка» есть текстовой блок «деловая проза» [Леонтьева, 1986]. В основном подкорпусе Национального корпуса русского языка (ruscorpora.ru) содержатся «нехудожественные тексты», которые можно выбирать, во-первых, по «сфере функционирования», – представлена в том числе «официально-деловая» сфера; во-вторых, по «типу текста» (представлены «деловые документы», «законодательные документы», «правовые документы», «судебные документы», «нотариальные документы», внутри каждого типа доступен поиск по жанрам); в-третьих, по «тематике текста» (представлены, в частности, тематические кластеры «администрация и управление» и «право»). Тексты официально-деловой сферы составляют 3,2% от объёма Основного корпуса, см. [Статистика корпуса].

Юридические документы, всего 441 текст, в основном – кодексы, входят в Открытый корпус (opencorpora.org). Некоторое количество деловых текстов (из-за отсутствия жанровой разметки трудно сказать – какое) попало и в состав Russian Business Corpus, одного из корпусов С. Шарова на сайте Университета Лидса и в другие русскоязычные веб-корпусы.

Основная часть отчёта о НИР

Описываемое междисциплинарное исследование развивалось по трём основным направлениям: I. Исследование объективной сложности, II. Исследование перцептивной трудности, III. Мониторинг сайтов гос. органов и исследование понятности (и правовых последствий непонимания юридических документов гражданами). В первом направлении работали корпусный лингвист и специалист по машинному обучению и обработке естественного языка, во втором – социолингвисты и антропологи, в третьем – юристы.

Обращение к языку правовых текстов в контексте изучения сложности, трудности и понятности неслучайно. Юридический язык (особенно – язык законов) имеет плохую репутацию и критикуется как сложный, тёмный, запутанный, для неюриста непонятный, ср. остроумную цитату из [Assy, 2011, 376]: “Complaints about the excessive complexity of the law are as old as the law itself” (жалобы на чрезмерную сложность закона стары, как сам закон).

I. Исследование объективной сложности правовых текстов

I.1. На протяжении трёх лет реализации проекта мы копили, обрабатывали и анализировали юридические текстовые коллекции. Нашей прикладной целю стало создание автоматической модели определения сложности русских правовых текстов, учитывающей значительное количество разнообразных (в том числе стилеспецифичных, то есть характерных для текстов официально-делового стиля) параметров.

Для её достижения на начальном этапе было необходимо: собрать текстовые коллекции, предобработать тексты, выбрать подходящие инструменты автоматического анализа текстов, разметить тексты. В результате создано три корпуса общим объёмом порядка 8 млн токенов. Принципиально важно, что наши три корпуса различаются в соответствии с типической фигурой адресата, на которого направлен конкретный юридический текст.

Первый корпус содержит локальные документы и называется “CorRIDA” Он включает документы, с которыми периодически сталкиваются носители языка-неюристы (формы информированных согласий, договоров, правил поведения в государственных учреждениях и пр., выкачанные с сайтов государственных учреждений). Корпус состоит из 1546 документов и содержит 1 млн 784 тыс. токенов.

Второй корпус содержит решения Конституционного Суда РФ и называется “CorDec”, он включает 584 документа, объём — 3 427 тыс. токенов. Решения пишутся высокопрофессиональными юристами и адресованы широкому кругу граждан, описание см. в [Blinova et al., 2020a].

Третий корпус содержит нормативные документы и называется “CorCodex”. Состав корпуса — 279 текстов кодексов, федеральных законов и постановлений правительства РФ (в общей сложности 3 млн 227 тыс. токенов). Адресатами таких текстов являются прежде всего профессиональные юристы.

I.2. Нам необходимо было выбрать инструменты автоматической разметки, позволяющие в дальнейшем успешно справиться с оценкой сложности коллекций (так как метрики, оценивающие значения конкретных параметров, учитывают далеко не только словоформы).

Для лемматизации, частеречной и синтаксической разметки мы использовали UDPipe. Показано, что синтаксические признаки хорошо предсказывают сложность текстов (и в целом было бы неверно оценивать сложность без обращения к синтаксическим

параметрам), см., например, [Ivanov et al. 2018]. Более того, корпусы, размеченные в формате UD (Universal Dependencies), в последнее время всё более активно используются при оценке морфосинтаксической сложности как при межъязыковом сопоставлении, так и при сравнений текстов (коллекций текстов) на одном языке, см., например, [Berdicevsks et al., 2018], [Çöltekin, Rama, 2018], [Yan, Kahane, 2018], [Dyer, 2018].

Отдельной задачей стал выбор между моделями UDPipe (существуют модели “ru-syntagrus”, “ru-gsd”, “ru-taiga”). Основанием для принятия решения стала статистика метрик, показывающая аккуратность работы моделей [Universal Dependencies 2.5 Models for UDPipe], [Straka, 2017]. Согласно этой статистике, где даны значения метрик аккуратности по параметрам UAS, LAS, MLAS и BLEX [CoNLL 2018 Shared Task], модель “russian-syntagrus-ud-2.5” работает лучше, поэтому нами выбрана именно она.

Как инструмент подробного морфологического анализа взят pymorphy2 [Korobov, 2015], так как частеречная разметка pymorphy2 позволяет, в частности, различать ‘ADJF’ – полные прилагательные, ‘ADJS’ – краткие прилагательные, ‘VERB’ – глаголы в личной форме, ‘INFN’ – инфинитивы, ‘PRTF’ – полные причастия, ‘PRTS’ – краткие причастия и ‘GRND’ – деепричастия. Это удобно для оценки сложности, в частности, потому, что выявлена положительная корреляция между количеством полных прилагательных (а также причастий и деепричастий) и сложностью и отрицательная корреляция между количеством глаголов в личной форме и сложностью, см. [Дружкин 2016].

I.3. Мы разметили три текстовые коллекции и получили корпусы. Наши юридические корпусы – это файлы в формате *json со слоями разметки, показанными в примере ниже, то есть:

- текст сегментируется на предложения,
- каждое предложение текста имеет номер,
- на слой "sentence" записывается предложение целиком,
- на слое "root" записана вершина,
- каждый токен предложения имеет номер,
- на слое "words" записан сам токен (словоформа или знак пунктуации),
- на слое лемм записана лемма, полученная UDPipe, на слое "pos" – тег части речи в терминах UDPipe,
- на слое "dep" записано синтаксическое отношение UDPipe,
- на слое "morph" – морфологическая разметка (часть речи + подробный морфологический разбор pymorphy2).


```
"866": {
"sentence": "Если совместное исполнение образует неразрывное целое, ни один из членов коллектива исполнителей не вправе без достаточных оснований запретить его использование.",
"root": "вправе",
"words": {
"0": {
"word": "Если",
"lemma": "если",
"pos": "SCONJ",
"dep": "mark",
```

"morph": "CONJ"

I.4. Для оценки сложности как текстовой переменной лингвисты используют различные языковые параметры (признаки, features) и метрики, с помощью которых эти признаки оцениваются. При этом все параметры вслед за [Tuldava, 2004] можно разделить на латентные (скрытые) и формально-статистические (поверхностные). Первые поддаются измерению, хотя и не поддаются непосредственному наблюдению в форме отдельных языковых сущностей, присутствующих в текстах на языке.

Соответственно, на этапе, следующем за сбором и разметкой данных (т.е. юридических корпусов), мы сконцентрировались на подборе параметров. Среди параметров есть скрытые – они взяты в основном с опорой на литературу о сложности (сюда же относятся формулы читабельности), и формально-статистические – они подбирались в основном с опорой на работы по стилистике и стилеметрии. Соответственно, в результате анализа литературы и анализа собранных данных мы получили набор параметров, которые, по нашему предположению, могут успешно диагностировать сложность (причём хорошо подходят для диагностики сложности юридических текстов, то есть способны работать в задачах классификации по сложности правовых текстов).

В результате мы получили 130 метрик, способных оценивать: скрытые от непосредственного наблюдения характеристики текстов, лексическую сложность, синтаксическую сложность, дискурсивную связность, а также (пока – в незначительной степени) неопределённость (vagueness), оказывающую влияние на понятность. Кроме того, введена одна метрика, оценивающая гипертекстовые связи внутри корпуса правовых текстов.

Следует заметить, что многие метрики, традиционно используемые для оценок сложности, не подтверждают свою действенность в оценке актуальной трудности в рамках психолингвистических экспериментов. Так, [Charrow, Charrow, 1979] выяснили, что длина предложения в стимуле практически не оказывала влияния на то, насколько успешно испытуемые справлялись с экспериментальным заданием. Они же показали, что диагностическая сила формул читабельности для оценки актуальной трудности невысока [Charrow, Charrow, 1979, 1341].

Следует заметить также, что и при интерпретации значений конкретных метрик в применение к правовым текстам могут возникать серьёзные разногласия. В частности, они касаются значений метрики TTR – отношение числа типов (уникальных словоформ или лемм документа) к числу токенов (всех словоформ или лемм документа). Среди нелингвистов, изучающих сложность русских юридических документов, существует противоположный традиционному подход к интерпретации значений TTR, цитируем: «множество формальных повторов одних и тех же слов, обозначающих субъектов права и различные юридические термины, мешают восприятию смысла предложения. В данном случае мы можем сказать, что уменьшение разнообразия не только не приводит к упрощению текста, но и вызывает обратный эффект» [Кучаков, Савельев 2018].

С тем, что в юридических текстах встречается множество повторов, трудно спорить, см. об этом, например, [Williams 2004, 113]. Однако представляется, что наличие повторов не только может утомлять читателя, но во многих случаях позволяет избежать проблем с интерпретацией языковых выражений, находящихся в кореферентных отношениях (то есть повторы являются средством референциальной связности). Повторы могут возникать из-за стремления избежать использования редуцированных референциальных средств

(местоимений), то есть боязни референциальной неоднозначности, а у авторов юридических текстов стремление к точности зачастую побеждает стремление к краткости. Кроме того, стоит помнить, что на процесс восприятия текста оказывает влияние эффект прайминга (в частности, лексического прайминга), а операции, связанные с разрешением кореферентных отношений, затрудняют обработку текста читающим или слушающим.

Добавим, что различные параметры оценки сложности слова не независимы друг от друга, в частности, согласно закону краткости Ципфа (Zipf's law of abbreviation), длина слова коррелирует с его частотой, см., например, [Bentz, Ferrer-i-Cancho, 2016]. Такое отношение между параметрами (из каких-то общих соображений) не очень хорошо влияет на успешность решения задач классификации.

Итак, разработан список метрик оценки сложности. При разработке последовательно применялись аналитические процедуры (введение каждой метрики обосновывалось).

1) В описываемой модели учитывается доля длинных словоформ и лемм (word_long_pr и lemma_long_pr). «Длинными» считаются слова и леммы длиной 4 слога и более.

2) В модели учитывается средняя длина словоформы в буквах (ACW) и слогах (ASW), а также среднее число букв на 100 словоформ (L). Средняя длина слов так или иначе используется в различных моделях оценки сложности и классификации текстов по сложности как один из многочисленных параметров, см., например, [Schwarm, Ostendorf 2005]. Значение L необходимо для вычисления индекса Колман-Лиау, но может быть информативным и само по себе.

3) Используются также значения параметров ASL (средняя длина предложения в словах) и ASS (средняя длина предложения в слогах), а также S (среднее число предложений на 100 словоформ). S используется для вычисления индекса Колман-Лиау.

4) Вычисляется доля запятых относительно всех токенов документа. По сути, это – простая метрика синтаксической сложности. Как указывает [Костенко, 2005], «что законодательный текст должен по возможности содержать короткие фразы <…>, а значит, он должен содержать как можно больше точек и меньше запятых».

5) Вычисляются индексы лексического разнообразия TTR (Type-Token Ratio), в качестве относительно более надёжных и независимых от длины текста мер используются индекс K (Yule's K) и индекс I (Yule's I).

6) Вычисляются доли гапаксов (hapax legomena – лемм, встречающихся в документе однократно, hapax dislegomena – лемм, встречающихся встречаются в документе дважды).

7) Вычисляются значения пяти адаптированных для русского формул читабельности.

8) Учитываются доли слов разных частеречных классов (частично – в терминах UD, частично – в терминах pymorphy2, список приведён в форме 1.4 настоящего отчёта).

Из литературы вопроса мы знаем, что в текстах ОДС по сравнению с другими стилями меняется частота вхождения слов разных частей речи. Сказанное подтверждается, например, в [Браславский, 2001], где наблюдается «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], [Дубовик, 2017], а также [Катинская, 2016] и др.

9) Учитываются n-граммы частеречных тегов (частеречная сочетаемость, список приведён в форме 1.4 настоящего отчёта.

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 генитивными аргументами.

В дескриптивной литературе по стилистике описываются, а в прескриптивной литературе – строго осуждаются так называемые «цепочки форм родительного падежа» (ср. также понятия «родительный приимённый», «нанизывания падежей»). В принципе, для оценки юридических текстов учитываемую длину последовательности 'NOUN + NOUN,*gent' можно было увеличить до трёх и более единиц. Так, исследование именных групп более чем с двумя генитивными аргументами на материале корпуса CorCodex (то есть текстов законов) показало, что максимальная наблюдаемая длина группы с ветвящимся генитивом достигает восьми элементов, см. [Веденина, 2021]. Приведём некоторые примеры из [Там же]: ‘компетенция органов государственной власти’, ‘установление порядка формирования доходов федерального бюджета’, ‘выравнивание уровня минимальной бюджетной обеспеченности субъектов Российской Федерации’, ‘определение основ исполнения бюджетов всех уровней бюджетной системы Российской Федерации’, ‘порядок согласования распределения расходов совместного ведения субъектов Российской Федерации’, ‘определение основ составления проектов бюджетов всех уровней бюджетной системы Российской Федерации, обеспечения необходимой степени конфиденциальности рассмотрения источников финансирования дефицита федерального бюджета’.

10) С помощью девяти метрик оцениваются доли лемм, принадлежащих разным частотным диапазонам Zipf value [0;8] (по созданному в рамках проекту сводному частотному списку лемм, собранному из частотных списков 4-х больших русских корпусов). Принятый подход позволяет тщательно учитывать вхождения в текст редких единиц, например, аббревиатур, см. пример со случайно выбранными леммами сводного частотного списка и значениями меры Ципфа [1;4], то есть приведены низкочастотные и среднечастотные леммы (все они отсутствуют в «Новом частотном словаре русской лексики»), см. Таблицу 1.

Таблица 1. Пример представления данных о частотности

лемма	**ipmArane um**	**ipm Taiga**	**ipm RuTenTe**	**ipm НКРЯ**	**Zipf Araneum**	**Zipf Taiga**	**Zipf RuTenTe**	**Zipf НКРЯ**	**Fcl araneum**	**Fcl taiga**	**Fcl rutenten**	**Fcl rnc**	**Zipf avrg**
ВКР	0,21	0,00	0,03	0,00	2	1	1	-	21	22	20	-	**1**
госнаркоконтроль	3,67	1,31	0,86	0,00	4	3	3	-	17	14	15	-	**3**
государственно-частный	25,6	0,73	1,89	0,00	4	3	3	-	14	15	14	-	**4**
единовластие	2,08	0,19	0,27	0,00	3	2	2	-	17	17	17	-	**3**
инновационность	7,88	0,11	0,73	0,00	4	2	3	-	15	17	16	-	**3**
концессионный	14,5	0,89	0,00	0,00	4	3	-	-	15	14	-	-	**4**
межэтажный	9,01	0,22	0,42	0,00	4	2	3	-	15	16	16	-	**3**
минкульт	9,14	0,25	0,35	0,00	4	2	3	-	15	16	17	-	**3**

мосэнерго	6,19	0,19	1,64	0,00	4	2	3	-	16	17	14	-	**3**
нравственно-религиозный	0,61	0,01	0,08	0,00	3	1	2	-	19	21	19	-	**2**
обезличивание	3,20	0,20	0,22	0,00	4	2	2	-	17	17	17	-	**3**
оцифровывать	9,53	0,45	0,83	0,00	4	3	3	-	15	15	15	-	**3**
санкционирование	3,23	0,07	0,27	0,00	4	2	2	-	17	18	17	-	**3**
телетрансляция	3,08	0,33	0,37	0,00	3	3	3	-	17	16	17	-	**3**
УФСБ	13,4	6,74	1,98	0,00	4	4	3	-	15	11	14	-	**4**
ФГОС	69,4	0,15	1,68	0,00	5	2	3	-	12	17	14	-	**3**

11) В модели использована одна словообразовательная метрика, с её помощью оцениваются доли лемм с рядом слоовообразовательных аффиксов (технически – заканчивающихся на буквенные последовательности ция, *ние, *вие, *тие, *ист, *изм и т.д.). На выбор метрики повлияло соображение, согласно которому производные слова в общем случае сложнее непроизводных.

12) Учитывается род существительных, так как абстрактные существительные (далеко не только учтённые в предыдущем пункте существительные на -ение, -ство и пр.) употребительные в ОДС, часто среднего рода.

13) Учитывается доля словоформ в родительном, творительном, именительном, дательном падеже.

14) Учитываются формы времени (финитных глаголов). О временных формах глагола из литературы известно, что в ОДС преобладают формы настоящего времени, при этом существует особое понятие «настоящее предписания». Более редким формам прошедшего времени в ОДС приписывается значение «подчёркнутой констатации». При этом употребление форм времени имеет жанровую специфику, поэтому может сыграть роль в оценке сложности. Можно думать, что жанры различаются по сложности (в некотором общем случае текст кодекса сложнее текста акта приёмки-передачи), то есть временная форма – это «сопутствующий» жанровой сложности признак.

15) Учитываются формы вида. В литературе по функциональной стилистике принято указывать резкие различия в употребительности глаголов несовершенного и совершенного вида разных жанрах. Соответственно, справедливо предположение, высказанное в пункте 13) выше относительно форм времени.

16) Учитывается лицо глагола. 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. Соответственно, справедливо предположение, высказанное в пунктах 13) и 14) выше относительно форм времени и вида.

17) Учитывается доля личных глагольных форм на -ся.

18) Тип конструкции (активная/пассивная) прямо или косвенно учтён в ряде метрик. В применении к категории «отдельные граммемы» это доля полных страдательных причастий (метрика «Pssv_prtf_pr») и доля кратких страдательных причастий (метрика «Pssv_prts_pr»).

19) Оценивается доля леммы ‘являться’. Эта метрика оценивает вхождения в текст предложений с составным именным сказуемым и по сути является синтаксической, а не собственно лексической.

20) Оценивается частотность лексических средств текстового дейксиса типа ‘(выше / ниже)названный’, ‘(выше / ниже)описанный’, ‘(выше / ниже)перечисленный’, ‘(выше / ниже) упомянутый’, ‘данный’ и пр. Такие единицы в модели оценки сложности задаются списком без «хвостов» (или финалей, список включает 29 стемм).

21) Учитывается доля графических сокращений, аббревиатур. Указанные единицы имеют отношение к анализу сложности, так как среди них много редких и сверхредких, то есть с высокой вероятностью незнакомых читателю, единиц. Кроме того, аббревиатуры могут быть непохожими на другие слова языка с точки зрения фонотактики, в них встречаются нетипичные для русского стечения согласных и гласных. Приведём пример из сводного частотного списка с указанием значения меры Ципфа (см. Таблицу 2). Видно, например, что хорошо знакомая всем работникам высшей школы аббревиатура «ВКР» обладает низкой общеязыковой частотностью.

Таблица 2. Данные о частотности аббревиатур (по сводному частотному списку)

лемма	ipm Araneum	ipm Taiga	ipm RuTenTen	ipm НКРЯ	Zipf Araneum	Zipf Taiga	Zipf RuTenTen	Zipf НКРЯ	Zipf avrg
ЦАО	20,33	0,81	3,79	0	4	3	4	-	4
ГДК	4,20	0,03	0,43	0	4	1	3	-	3
РСН	3,63	3,23	0,44	0	4	4	3	-	3
ОСО	4,85	0,59	1,32	0	4	3	3	-	3
ТЭО	9,75	0,10	1,26	0	4	2	3	-	3
ФАП	12,26	0,21	1,21	0	4	2	3	-	3
КТД	1,65	0,00	0,15	0	3	1	2	-	2
ПНК	1,48	0,03	0,11	0	3	2	2	-	2
ГЭФ	2,85	0,00	0,19	0	3	0	2	-	2
ДЭО	0,45	0,01	0,05	0	3	1	2	-	2
ВНУ	0,95	0,01	0,21	0	3	1	2	-	2
КСР	1,62	0,00	0,22	0	3	1	2	-	2
НКЛ	0,22	0,15	0,04	0	2	2	2	-	2
АЦО	0,09	0,01	0,02	0	2	1	1	-	1
УМД	0,24	0,01	0,02	0	2	1	1	-	1
КЗЦ	0,23	0,01	0,01	0	2	1	1	-	1
ТДО	0,20	0,00	0,03	0	2	1	1	-	1
ЮПК	0,10	0,01	0,02	0	2	1	1	-	1
КЗМ	0,14	0,00	0,02	0	2	1	1	-	1
ВКР	0,21	0,00	0,03	0	2	1	1	-	1

22) Для описания гипертекстовых связей (когда для понимания какого-то положения закона необходимо обратиться к тексту другого закона) введена метрика «FZ_pr».

23) Учитываются доли юридических терминов (термины заданы списком).

Юридические термины служат признаками сложности, так как для их понимания требуются специальные знания. Можно заметить также, что они зачастую являются заимствованиями (ср. *диспаша*, *диспонент*, *дистрибьютор*, *дистрикт*) и в некотором общем случае характеризуются низкой общеязыковой частотностью, см. некоторые иллюстрации ниже, где даны первые 10 позиций алфавитного списка юридических терминов с оценкой частотности по сводному частотному списку. Частотность «0,00» значит, что лемма в списке есть, но значение меры Ципфа при округлении до целых равно нулю. Прочерк значит, что леммы в списке нет.

Таблица 3. Данные о частотности терминов (по сводному частотному списку)

лемма	**ipm Araneum**	**ipm Taiga**	**ipm RuTenTen**	**ipm НКРЯ**	**Zipf Araneum**	**Zipf Taiga**	**Zipf RuTenTen**	**Zipf НКРЯ**	**Zipf avrg**
абандон	0,37	0,00	0,02	0.0	3,00	-	1,00	-	2
абдикация	-	-	-	-	-	-	-	-	-
абекор	-	-	-	-	-	-	-	-	-
аболиционизм	0,20	0,01	0,02	0,00	2	1	1	-	1
аболиция	0,03	0,00	0,00	0,00	1	0,00	-	-	1
абонемент	77,67	1,86	5,97	1,8	5	3	4	3	4
абонент	264,13	9,08	43,80	10,6	5	4	5	4	5
абориген	39,23	2,35	4,95	5,8	5	3	4	4	4
аборт	123,14	5,41	12,72	8,3	5	4	4	4	4
аброгация	-	-	-	-	-	-	-	-	-

24) Учитывается доля абстрактных лемм (относительно всех лемм текста).

25) Оцениваются вхождения однословных лексических показателей деонтической возможности и необходимости. Список показателей получен в результате анализа частотного списка трёх юридических корпусов, созданных в рамках проекта.

26) Учитываются доли неоднословных союзов и предлогов. Использовать информацию о вхождениях в текст неоднословных единиц при оценке сложности – осмысленный шаг уже потому, что мысленная операция замены однословного выражения на неоднословное увеличивает длину высказывания. В то же время длина предложения сама по себе (по данным психолингвистических исследований трудности) – ненадёжный показатель, поэтому проверять актуальную трудность семантически эквивалентных длинных и коротких языковых выражений нужно в эксперименте.

27) Учитывается доля конструкций с лёгкими глаголами.

28) В списке признаков слоя разметки “dep” присутствуют:

- признаки, показывающие организацию отдельных синтаксических групп (именной группы – «Amod_p», доля адъективных модификаторов имени, глагольной группы – «Advmod_pr», доля наречных модификаторов предиката);

- признак, описывающий вхождения аппозитивных именных групп («Appos»);

- признаки, показывающих наличие сочинённых рядов (будь то клаузы или однородные члены предложения; имеется 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» отдельно от 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»).

29) Кроме того, в списке признаков слоя разметки “dep” присутствуют дискурсивные метрики («Parataxis», «Discourse», см. форму 1.4 настоящего отчёта).

30) Наконец, двумя метриками дополнительно оценивается связность (а именно, референциальная связность и повторяемость граммем времени и вида в личных формах глагола в смежных предложениях).

I.5. Получены результаты тестирование модели оценки сложности.

I.5.1. Выполнено тестирование на текстовом наборе “plainrussian” И. Бегтина.

I.5.2. Выполнено тестирование на текстовом наборе учебников обществознания.

I.5.3. Выполнена классификация с использованием в качестве параметров векторов языковой модели USE (Universal Sentence Encoder).

Итоговые показатели качества таковы:

- для кодирования с использованием метрик (130 параметров, plainrussian): средняя точность – 88% со среднеквадратичным отклонением 9%;

- для кодирования с использованием метрик (130 параметров, учебники): средняя точность – 90% со среднеквадратичным отклонением 5%;

- для кодирования с использованием языковой модели (768 параметров): средняя точность – 70% со среднеквадратичным отклонением 15%.

Вывод таков: наши метрики позволяют получить более точные и согласованные оценки сложности текстов, чем классификация с параметрами USE.

I.6. Получены оценки отдельных метрик

В тестировании получены данные об эффективности работы 130 метрик в задаче классификации по сложности. Важно заметить, что тестирование проводилось на наборах данных, существенно отличающихся от наших. Между тем, метрики были целенаправленно разработаны прежде всего для применения к юридическим текстам (текстам ОДС). Сказанное значит, что в текстах других стилей некоторые учитываемые признаки могут описывать редкие или сверхредкие явления.

В целом результаты оценки работы отдельных метрик таковы. Эксперимент с текстовым набором “plainrussian” показал, что для задачи классификации значимы 72 метрики. Эксперимент с учебниками обществознания показал, 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 (в разной степени) 94 других метрики. В наборах работающих на классификацию по сложности признаков совпадает 57 признаков.

В задачах классификации хорошо сработали формулы читабельности. Это можно было бы объяснить тем, что формулы действительно хорошо справляются с предсказанием сложности. Однако, как уже было сказано выше, текстовые наборы, на которых мы проверяли эффективность метрик, – это и есть наборы, на которых используемые формулы читабельности разрабатывались.

I.7. На сайте проекта опубликованы три юридических корпуса и таблицы с метаданными, где содержатся значения 130 метрик сложности, присвоенных документам корпусов.

I.8. Сформированы частотные списки лемм трёх корпусов со значениями абсолютных и относительных частот, мер дисперсии D Жуйана и DP Гриса. Подготовлен общий частотный список по трём корпусам (в том же формате). Данные опубликованы на сайте проекта.

I.9. Подготовлен сводный частотный список лемм с условным названием «Фреквентатор» (примерно 1 млн строк) с информацией о распределении лемм по зонам частотного словаря (для распределения использовались две меры частотности – Zipf Value и FClass). Список получен по данным четырёх больших русских корпусов.

I.10. Разработана архитектура гибридной модели оценки сложности, в которой на классификацию по сложности работают и наши метрики, и результаты кодирования текстов языковой моделью USE.

II. Исследование перцептивной трудности правовых текстов

Материалом «перцептивного» направления проекта стал опрос граждан РФ, проведенный в 2018–2020 гг., целью которого было узнать, в какой степени люди (различного возраста, с тем или иным образованием и опытом работы) способны понять язык официальных документов, и что составляет для них трудности при восприятии. Участникам опроса предлагалось ответить на ряд вопросов по трём текстам, среди которых:

- информированное согласие на медицинское вмешательство (приводилось полностью),
- правила поведения в заповеднике (полностью),
- правила приема в вуз (в опросе представлен фрагмент).

Соответствующие части опроса в дальнейшем обозначаются как Анкета 1 («Информированное согласие» или «Медицина»), Анкета 2 («Правила поведения» или «Культура»), Анкета 3 («Правила приема» или «Образование»).

Исследователи, изучавшие трудность русских официальных документов в ходе опроса, получили следующие основные результаты.

II.1. Получен актуальный на момент старта проекта список респондентов с распределением по группам в соответствии с возрастом, образованием, профессией и занятиями, а также в соответствии с компетенциями в работе с документами.

II.2. Собраны заполненные респондентами электронные и бумажные анкеты. Все данные о респондентах представлены в таблице, содержащей поля Name, Age_Group, Educ, Prof, Experience и Gender.

II.3. Разработана система обозначения типовых ответов и выставления баллов, отражающих компетенции респондентов в работе со сложными текстами.

II.4. Завершено создание реляционной базы данных «Questionnaire», содержащей информацию о 400 респондентах массового социолингвистического опроса (т.е. таблицу «Respondents» с заполненными полями Name, Age_Group, Educ, Prof, Experience, Gender), а также включающая ответы на все три части опроса (т.е. таблицу «Answers» с заполненными полями, содержащими коды для типизированных ответов на вопросы первой, второй и третьей части анкеты).

II.5. Проведён анализ ответов респондентов по доменам (образование, культура, медицина), а также анализ индивидуальных и групповых различий компетенций в работе со сложными текстами с точки зрения объективных факторов формирования конкретного вида компетенции.

В частности, детальное рассмотрение ответов респондентов на анкету домена «Культура» показало следующее.

В опрос были включены вопросы разного типа:

1. вопросы по тексту, направленные на проверку понимания документа;
2. вопросы о языковых особенностях текстов, которые вызывают затруднения при чтении;
3. вопросы, направленные на проверку понимания более общих характеристик текста, как-то: его юридического смысла, организации информации, которая определяется более широким контекстом (например, о ходе лечения и возможности его изменения);
4. вопросы оценочного характера, выявляющие отношение носителей языка к документу в целом и его конкретным языковым особенностям;
5. вопросы, проверяющие отношение к правилам поведения в общественных местах и запретам.

По вопросам 5 типа выявлялась информация об отношении участников опроса к языку документа. Оценочные эпитеты, которые респонденты дали общему тону документа и правилам (вопросы 13 и 14), разбиваются на негативные, нейтральные и положительные (при этом они составляют континуум и в некоторых случаях представляют пограничные случаи). Выяснено, что в ответах представлен весь спектр оценок, например: абсолютно неуважительный; тон старой злой советской вахтерши; хамско-командный; адекватно-нейтральный, официальный, предупредительный; вежливый, доходчивый; положительный.

Предварительный анализ полученных ответов позволяет выделить определенные типажи респондентов на основании их отношения к запретам и языку официальных документов:

- считают обязательным соблюдение правил и считают, что правила необходимо оформлять в строгой форме;
- критически относятся к правилам, не считают обязательным их соблюдение, предпочитают краткие перечни в некатегоричной, вежливой форме).

II.6. Для выявления индивидуальных и групповых различий компетенций в работе со сложными текстами с точки зрения объективных факторов формирования конкретного вида компетенции оценена индивидуальная компетенция респондентов, а также средняя компетенция по группе, в зависимости от уровня образования, возраста, опыта работы с документами (нет опыта/ небольшой/ значительный), занятости в последние 5-10 лет (связана со сложными текстами/ не связана).

II.7. Сформирован список перцептивно трудных собственно языковых явлений, выделенный по результатам анализа ответов респондентов.

1) Трудности первого порядка вызывает незнакомая лексика, в том числе (относительно) употребительные термины, например, «антибиотики» и «анальгетики»; «аллергические реакции»; «хирургическое вмешательство»; «зубной протез»;

2) трудности второго порядка связаны с такими особенностями, как распространенные предложения, многокомпонентные слова и устойчивые обороты, относящиеся к официальному стилю, синтаксические шаблоны (напр., последовательности генитивов) – тем не менее, доступными для интерпретации содержания при условии знакомства с фактической ситуацией, описываемой в документе;

3) трудности третьего порядка вызваны строго официальным или официально-юридическим типом текста, с характерными признаками стилевой маркированности в грамматическом оформлении, лексемами обобщенного характера, с осложнениями, в основном связанными с длиной слов, словосочетаний, фраз и текста в целом. Эти особенности характеризуют текст «Правил приема» (домен «Образование»);

4) трудности четвертого порядка связаны с введением новой терминологии на базе общеупотребительной лексики – п. 9 и п. 10 «Правил приема», где вводится значение словосочетаний «основания приема» и «правила поступления».

Так, «Правила приема» вызывают крайне негативную реакцию респондентов, а кроме того, их комментарии эксплицитно свидетельствуют о возникающих трудностях, напр.: «Слова в предложениях, из которых состоит текст, кажутся несогласованными между собой. Суть трудно понять» (реставратор, среднее специальное образование); «Текст очень сложен (особенно для детей, окончивших школу). Неудобочитаемый. Очень сложны и недоступны передаваемые смыслы. Переформулировать фразы можно в каждом положении» (разнорабочий, студент).

II.8. Проведен анализ интервью с респондентами с целью выявления их отношения к признакам канцелярита в административных объявлениях.

II.9. Выполнена оценка ответов респондентов в соответствии с базовыми социальными параметрами (возрастом, образованием, полом), а также опытом работы с документами и типом занятости.

II.10. Подробно исследовано отношение к языку официальных документов, см. форму 1.4 настоящего отчёта.

II.11. Выявлено наличие зависимости между проявленным отношением к представленному в Правилах поведения (Анкета-2) языковому варианту и социальными параметрами респондентов, см. форму 1.4 настоящего отчёта.

II.12. Получены результаты исследования функционирование официального медицинского документа (на материале опроса и на материале полуструктурированных интервью с врачами, проведённых на этапе 2020 г.), см. форму 1.4 настоящего отчёта.

II.13. Получены данные о субъективной сложности (трудности) исследуемых текстов, анализ позволил судить о субъективной сложности текстов и вопросов на основании того, какие именно вопросы вызвали больше всего затруднений у респондентов. Количество отказов по Анкете-3 было самым большим – 25,2% от общего числа ответов.

II.14. Описаны стратегии ответов респондентов, объединённых в группы по параметрам возраст, пол, образование, опыт работа с документами и тип занятости.

III. Мониторинг сайтов гос. органов и исследование понятности (и правовых последствий непонимания юридических документов гражданами)

Выполнен мониторинг 371 сайта государственных органов. Мониторинг позволил оценить соответствие содержания официальных сайтов требованиям доступности информации о деятельности государственных органов и органов местного самоуправления, а также рассмотреть информационное содержание официальных сайтов с точки зрения его понятности адресату – пользователю сайта. Для решения этой задачи введены критерии рассмотрения сайтов: во-первых, полнота информации (содержательный критерий), во-вторых, наличие/отсутствие юридико-лингвистической неопределенности, в-третьих, уровень читабельность текстов на официальных сайтах (формально-языковой критерий).

В рамках мониторинга сделано прежде всего следующее.

III.1. Выработана методика проведения мониторинга официальных сайтов государственных органов и органов местного самоуправления:

а) разработаны критерии включения сайтов в список сайтов, подлежащих мониторингу;

б) определён перечень сайтов для мониторинга.

в) сформулированы критерии оценки языкового контента сайтов, написана «Инструкция по проведению мониторинга официальных сайтов государственных органов».

III.2. С использованием подготовленного на этапе 2019 г. метода, инструкции и списка сайтов, подлежащих мониторингу, выполнен первый этап мониторинга официальных сайтов государственных органов и органов местного самоуправления (подразумевающий мониторинг сайтов судов). Конкретнее, проведен мониторинг 205 официальных сайтов судов общей юрисдикции и арбитражных судов Северо-Западного федерального округа РФ.

III.3. Проанализированы параметры соответствия информационного наполнения официального сайта о деятельности судов в РФ требованиям нормативно-правовых актов.

Показатели параметра – информация, наличие которой оценивается, – разработаны с учетом требований нормативно-правовых актов, регулирующих деятельность государственных органов, органов местного самоуправления по обеспечению доступа к информации о деятельности государственных органов, органов местного самоуправления.

Отдельно выделены показатели параметра для судов, поскольку перечень информации о деятельности судов в Российской Федерации устанавливается Федеральным законом "Об обеспечении доступа к информации о деятельности судов в Российской Федерации".

В зависимости от наличия и полноты сведений на официальном сайте, каждому сайту присваивалось числовое значение в соответствии со следующим принципом:

0 – информация отсутствует,

1 – информация представлена, но не в полном объеме,

2 – информация представлена в полном объеме.

Конкретные признаки, используемые для анализа информационного наполнения сайтов показаны в файле с дополнительными материалами. В общей сложности выделено информационных 27 признаков. Среди них, например, представленность следующей информации: «Полномочия суда», «Перечень законов, регламентирующих деятельность суда», «Регламент суда, инструкция по делопроизводству в суде и иные акты,

регулирующие вопросы внутренней деятельности суда», «Порядок обжалования судебных актов» и мн. др.

III.4. Проанализировано наличие юридико-лингвистической неопределенности в текстах официальных сайтов.

Параметр отражает наличие либо отсутствие неопределенности в текстах официального сайта. Под юридико-лингвистической неопределенностью для целей мониторинга понимается любая неопределенность, которая не позволяет однозначно установить содержание текста и порождает произвольное понимание того, что означают положения по существу. О юридико-лингвистической неопределенности свидетельствуют использование лексики, недопустимой для деловых текстов, употребление в тексте выражений типа «обычно», «как правило», «чаще всего», «по мере необходимости», «в исключительных случаях», наличие в тексте ошибочного употребления синтаксических конструкций, употребление синтаксических конструкций, затрудняющих восприятие смысла.

В зависимости от наличия юридико-лингвистической неопределенности на официальном сайте каждому сайте присваивалось одно из трех значений критерия:

0 – наличие юридико-лингвистической неопределенности свыше 3 раз встречается на официальном сайте;

1 – наличие юридико-лингвистической неопределенности от 1 до 3 раз встречается на официальном сайте;

2 – юридико-лингвистическая неопределенность отсутствует.

III.5. Проанализировано текстовое содержимое сайтов с использованием формул читабельности.

III.6. Рассчитаны коэффициенты итогового показателя доступности и открытости официального сайта (КД).

Оценён итоговый показатель доступности и открытости сайтов, который может составлять от 0 до 100, при этом чем выше коэффициент, тем тексты на официальном сайте более доступны и открыты. Для получения итогового показателя для каждого сайта использовалась формула KZ – b, где b – индекс читабельности, полученный с использованием формул читабельности.

Расчёт и интерпретация значений коэффициентов выполнялись по следующей схеме.

Максимальное значение каждого коэффициента – 100. Оценка официального сайта происходила на основании совокупности коэффициентов отдельных параметров (KX – индекс соответствия информационного наполнения официального сайта о деятельности судов в РФ требованиям нормативно-правовых актов; KY – индекс наличия юридико-лингвистической неопределенности; KZ – итоговый показатель читабельности) и рассчитывалась по формуле:

(KX+KY+KZ)/300× 100= КД.

III.7. Создан рейтинг доступности и открытости официальных сайтов судов. Рейтинг сформирован путем сортировки официальных сайтов по убыванию значения коэффициента итогового показателя доступности и открытости КД.

III.8. По разработанной схеме завершён мониторинг официальных сайтов государственных органов и органов местного самоуправления (на этапе 2021 г. осуществлён мониторинг 166 сайтов, в общей сложности выполнен мониторинг 371 сайта).

IV. Подготовлена «Концепции использования государственного языка в деятельности государственных и муниципальных органов и организаций», см. дополнительные материалы.

Структура включает следующие основные разделы: «I. Общие положения», «II. Основные принципы использования государственного языка», «III. Методические рекомендации».

Концепция учитывает разнообразие видов правовых актов, особенности их коммуникативных характеристик, а также полученные в рамках других направлений настоящего исследования результаты выявления проблем понятности официальных документов и причин их возникновения.

V. Результаты проекта отражены в подготовленной к печати коллективной монографии.

Заключение

В рамках проекта создано **три юридических корпуса**. Разработана **модель оценки сложности русских правовых текстов**, основанная на 130 метриках. Предусмотрена оценка т. наз. скрытых параметров текста, она производится при помощи базовых метрик, традиционно используемых в задачах классификации по сложности (среди них: средняя длина предложения в словах, средняя длина слова в слогах, метрики лексического разнообразия, доля гапаксов и т. п. – всего 28 метрик); к области оценки скрытых периметров текста можно причислить и формулы читабельности (в модели используется 5 формул читабельности, ранее адаптированных для русского другими исследователями). Учтены и поверхностные параметры. Они разделены на категории «учёт слов разных частеречных классов», «n-граммы частеречных тегов (частеречная сочетаемость)», «частотность лемм», «словообразование», «отдельные граммемы», «лексические и семантические признаки, неоднословные выражения», «синтаксические признаки», «оценки связности».

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. Соответственно, наша модель оценки сложности способна учитывать доли лемм, принадлежащих каждой из 9-ти частотных зон и различать высокочастотные, среднечастотные и низкочастотные леммы.

Для обеспечения работы метрик «лексические и семантические признаки, неоднословные выражения» создан ряд пользовательских словарей и списков, среди которых:

- список лексических средств текстового дейксиса типа *‘(выше / ниже)названный’, ‘(выше / ниже)описанный’, ‘(выше / ниже)перечисленный’, ‘(выше / ниже) упомянутый’*,
- список графических сокращений (1,5 тыс. единиц) и аббревиатур (2 тыс. единиц), список юридических терминов (10 тыс. однословных и неоднословных терминов),
- список абстрактных лемм (17 тыс. единиц),
- список однословных лексических показателей деонтической возможности и необходимости, ср.: *‘дозволить’, ‘должен’, ‘допустимо’, ‘запрещать’, ‘можно’,*

‘надлежащий’, ‘неподобающий’, ‘неправомерно’, ‘приемлемо’, ‘противоправный’, ‘следовать’ (‘следует’) и т. д.,

- список конструкций с лёгкими глаголами (учтено 6 тыс. уникальных последовательностей лемм с пунктуацией типа *‘оказывать содействие’, ‘давать оценка’, ‘осуществлять свой деятельность’, ‘принять решение’, ‘осуществлять подготовка , переподготовка’*).

Важным компонентом модели является учёт признаков, основанных на разметке UDPipe (22 метрики).

Значения 130 метрик сложности подсчитаны для документов собранных и размеченных юридических корпусов. Это корпус законов CorCodex, корпус решений конституционного суда CorDec и корпус локальных актов CorRIDA (всего порядка 8 млн токенов, для разметки использованы UDPipe и pymorphy2). Принципиально важно, что три корпуса различаются в соответствии с типической фигурой адресата, на которого направлен конкретный юридический текст (CorRIDA включает документы, с которыми периодически сталкиваются носители языка-неюристы, а тексты CorCodex читают прежде всего профессиональные юристы).

Таким образом, получен набор юридических текстов с разметкой и оценкой сложности.

Модель оценки сложности протестирована, выявлены метрики, наиболее эффективно работающие на решение задачи классификации по сложности.

Предложенный ранее тестовый подход демонстрирует возможность применения метрик в качестве параметров для классификации текстов по сложности. Однако результаты, полученные для языковой модели USE (Universal Sentence Encoder), не дают полного представления об эффективности подобного подхода. В первую очередь это связано с общей структурой кодировщика, созданного для кодирования 18 различных языков для множества различных задач анализа естественного языка. Использование моделей, предобученных для работы с русским языком с последующим тюнингом для задачи определения сложности позволит точнее определить возможности подобных моделей в решении задачи классификации по сложности.

Таким образом, переходя к перспективам исследования, можно сказать, что для повышения эффективности работы модели можно, во-первых, добавлять языковые метрики, во-вторых, сделать модель гибридной, добавив в неё нейросетевую кодировку (которую затем использовать как отдельный параметр классификации по сложности).

Публикации

1. Белов С.А. Роль языка в обеспечении понятности и определенности нормативных правовых актов // Вестник Санкт-Петербургского университета. Право. 2022. Т. 2 (в печати, WoS, Scopus, РИНЦ).
2. Белов С.А., Кропачев Н.М. Представления сотрудников СМИ об источниках кодификации языковых норм и правил употребления языка (по материалам анкетирования) // Вестник Санкт-Петербургского университета. Язык и литература. 2021, Том 18, № 3. С. 512–527. URL: https://languagejournal.spbu.ru/article/view/12134/8369. doi 10.21638/spbu09.2021.306 (Scopus, РИНЦ)
3. Белов С.А., Ревазова Е.А., Руднева Е.А. Правовые нормы vs. представления врачей и пациентов о содержании и значении информированного согласия на медицинское вмешательство. Вестник Санкт-Петербургского университета. Право (2021). Т. 4 (в печати, WoS, Scopus, РИНЦ).
4. Блинова О.В., Алексеева Ю.Е. Личное местоимение как редуцированное референциальное средство в русском правовом тексте // Вопросы русского языка в юридических делах и процедурах. Международная научно-практическая конференция. – СПб.: Первый класс, 2021. С. 146-159 (в печати, РИНЦ).
5. Блинова О.В., Тарасов Н.В. Сложность русских правовых текстов: методы оценки и языковые данные // Труды международной конференции «Корпусная лингвистика-2021». — СПб.: Скифия-принт, 2021. С. 175-182. ISBN 978-5-98620-557-1 (РИНЦ).
6. Глазанова Е.В. Студентка vs водитель в возрасте, или почему никто не любит читать официальные документы // Социо- и психолингвистические исследования, Вып. 9, 2021. С.76-85. URL: http://splr.psu.ru/wp-content/uploads/2021/12/Glazanova.pdf (РИНЦ)
7. Гулида В., Руднева Е. «Сложно и глупо» vs «профессионально» и «вежливо»: отношение носителей русского языка к канцеляриту в объявлениях // Антропологический форум. 2021. № 50. С. 200–224. URL: https://anthropologie.kunstkamera.ru/files/pdf/050/gulida_rudneva.pdf, doi: 10.31250/1815-8870-2021-17-50-200-224 (Scopus, РИНЦ)
8. Руднева Е.А., Глазанова Е.В. Как россияне понимают текст информированного согласия на медицинское вмешательство // Материалы международной научно-практической конференции «V Фирсовские чтения: Современные языки, коммуникация и миграция в условиях глобализации» (в печати, РИНЦ).
9. Руднева Е.А., Гулида В.Б., Глазанова Е.В. Понимание российскими гражданами официальных документов (по результатам анкетирования) // Вопросы русского языка в юридических делах и процедурах. Международная научно-практическая конференция. – СПб.: Первый класс, 2021. С. 248-260 (в печати, РИНЦ).
10. Gulida V., Glazanova E., Rudneva E. How Russian Speakers Comprehend Documents // 16th Conference on Legal Translation and Interpreting and Comparative Legilinguistics (Legal Linguistics)/ Literature, Media and Law. Book of Abstracts. Coordinators: Botezat, Onorina; Le, Cheng Matulewska, Aleksandra. URL: http://limbi-straine.ucdc.ro/doc/Book%20of%20Abstracts_Legal%20Translation%20and....pd ISBN 978-606-26-1403-4

Библиография

Bentz C. & Ferrer-i-Cancho R. Zipf's law of abbreviation as a language universal // Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics (2016). University of Tubingen. URL: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68639 (date of access: 01.12.2021).

Charrow, R.P., Charrow, V.R. Making Legal Language Understandable: A Psycholinguistic Study of Jury Instructions // Columbia Law Review. 1979. Vol. 79, № 7. P. 1306-1374.

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. Vol. 17, P. 277-287.

Owens R. J., Wedeking J. P. Justices and legal clarity: Analyzing the complexity of US Supreme Court opinions // Law & Society Review. 2011. Vol. 45, № 4. P. 1027–1061.

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

Блинова О. В., Белов, С.А. Языковая неоднозначность и неопределённость в русских правовых текстах. Вестник Санкт-Петербургского университета. Право. 2020. Т. 11, № 4. С. 774-812.

Браславский П. Морфологический строй функциональных стилей (на материале документов Internet) // Известия Уральского государственного университета. 2001. № 21. С. 9-7.

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

Голуб И.Б. Стилистика русского языка. М., 2001.

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

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

Дубовик А. Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам: выпускная квалификационная работа магистра. СПб., 2017.

Ингве В. Гипотеза глубины // Новое в лингвистике. 1965. Вып. 4. С. 126-138.

Катинская А. Ю. Применение Многомерного анализа к изучению языковой вариативности в русскоязычных Интернет-жанрах // Материалы конференции «Диалог». URL: http://www.dialog-21.ru/media/3474/katinskaya.pdf (дата обращения: 01.12.2021)

Кнутов А. В., Плаксин С. М. и др. Сложность российских законов. Опыт синтаксического анализа. М., 2020.

Костенко М.А. Правовая лингвистика в законотворческом процессе // Известия ЮФУ. Технические науки. 2005. №9 (53). С. 127.

Кучаков Р.К., Савельев Д.А. Решения арбитражных судов субъектов Российской Федерации: лексическое и синтаксическое качество текстов: аналитическая записка / Под ред. Д. Скугаревского. СПб., 2019.

Кучаков Р.К., Савельев Д.А. Сложность правовых актов в России: Лексическое и синтаксическое качество текстов / Под ред. Д. Скугаревского. СПб., 2018.

Леонтьева Н.Н. Об информационной системе словарей Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. М., 1986. С. 109-125.

Мартыненко Г.Я. Статистическое исследование синтаксической сложности предложения (на материале болг. яз.) // Информ. вопросы семиотики, лингвистики и автомат. пер., 1971. Вып. 1. С. 84-101.

Оборнева И. В. Автоматизация оценки качества восприятия текста // Вестник Московского городского педагогического университета. 2005. № 2. С. 221-233.

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. С.347-356.

Савельев Д. А. О создании и перспективах использования корпуса текстов российских правовых актов как набора открытых данных // Право. Журнал Высшей школы экономики. 2018. № 1. С. 26–44.

Приложение

Концепция использования государственного языка в деятельности государственных и муниципальных органов и организаций

I. Общие положения

1. Настоящая Концепция описывает требования к использованию государственного языка государственными и муниципальными органами и организациями. Соблюдение этих требований обеспечит право граждан на информацию о деятельности указанных органов и организаций, а также о правах, обязанностях и ответственности граждан.

2. Право на информацию включает право на получение существенной, достаточной, своевременной, всесторонней, качественной и легко интерпретируемой информации, затрагивающей права и законные интересы граждан. Государственный язык обеспечивает политическое, экономическое, культурное, информационное единство страны, создает возможности для эффективной коммуникации в обществе, в том числе между частными лицами и органами публичной власти.

II. Основные принципы использования государственного языка

3. При составлении официальных документов, а также при публичном размещении информации государственные и муниципальные органы должны четко обозначить, кому адресован текст, и составлять его, учитывая предполагаемые языковые компетенции адресата.

4. В отношении каждого текста должно быть четко определено его содержание. Государственные и муниципальные органы и организации издают документы, которые:

1) устанавливают общеобязательные правила общего характера (нормативные акты),

2) в том числе в конкретном органе или организации (локальные нормативные акты),

3) содержат общеобязательные предписания для конкретной ситуации (общие административные акты),

4) разъясняют общеобязательные правила, установленные нормативными актами (разъяснения нормативного акта),

5) содержат конкретные предписания, адресованные конкретному лицу в конкретной ситуации (индивидуальные акты),

6) фиксируют фактические обстоятельства,

7) носят информационный характер и не содержат обязательных правил либо предписаний, а также не фиксируют факты.

В одном документе не должны совмещаться разные по содержанию положения (нормативные и индивидуальные предписания, правовые предписания и другая информация). Этот характер содержания документа должен быть четко обозначен до его составления и ясно выражен в тексте документа, ему должны соответствовать используемые языковые средства (с точки зрения использования модальных конструкций и т.п. – см. п. 11 настоящего документа). В тексте обязательно указывается, в какой части документ воспроизводит предписания других документов, а в какой части устанавливает новые.

5. При составлении любых текстов должны соблюдаться официально утвержденные правила орфографии и пунктуации, а также иные нормы современного русского

литературного языка, зафиксированные официально в качестве норм, обязательных для соблюдения при использовании языка как государственного.
6. При составлении официальных документов, а также при публичном размещении информации следует избегать неопределенности, двусмысленности и неоднозначности, утверждения должны носить категоричный, а не возможный либо предполагаемый характер.
7. Любой текст должен быть максимально краток – настолько, насколько это возможно при изложении всего необходимого его содержания.
8. При составлении любого текста следует использовать общеупотребительные слова (избегать специальной терминологии, диалектов и жаргонов) и максимально простые синтаксические конструкции, и одновременно соблюдать требования официального стиля (не допускать разговорных слов и выражений, эмоциональных или не нейтральных стилистически слов и речевых оборотов).
9. При организации информации в тексте должно обеспечиваться удобство ее использования, в частности, по возможности, структура должна соответствовать ситуациям, в которых адресату потребуется использовать содержание документа, должно даваться пошаговое описание необходимых действий и их четкое обозначение (кто, в какие сроки и каким образом что именно должен сделать). Текст должен разбиваться на максимально короткие разделы и подразделы с краткими содержательными заголовками. Вся существенная информация, необходимая для выполнения предписаний этого документа, должна быть изложена непосредственно в его тексте, без отсылок к другим документам.

III. Методические рекомендации

10. В отношении предполагаемого адресата документа или информации следует предполагать возраст, уровень образования, наличие или отсутствие профессиональных знаний. В случае, если круг адресатов предполагается широким и охватывает лиц, имеющих разные языковые и внеязыковые компетенции, то текст должен ориентироваться на тех, кто обладает минимальными знаниями и уровнем подготовки.
11. В отношении каждого документа из используемых слов и грамматических конструкций должно быть четко понятно, кто его адресат, какова его цель и содержание этого документа. В частности, нормативные предписания не должны выглядеть как констатация (описание) фактов; предписание должно отличаться от дозволения, прежде всего с помощью модальных глаголов (вместо «заявитель подает обращение» следует писать «заявитель должен подать обращение»). Дефиниции не должны содержать норм (предписаний или запретов).
12. В каждом предложении любого текста следует использовать четкую синтаксическую структуру. Рекомендуется использовать прямой порядок слов. В одном предложении не следует использовать более одного придаточного предложения. Допустимо перечисление однородных членов, оформленных в виде списка. Не следует использовать сложносочинённые предложения. Следует сперва излагать более общие, затем – более специальные положения, в том числе сперва общие правила, и только потом исключения.
13. В предложениях, содержащих правовые предписания, следует избегать использования скобок и знаков препинания, кроме запятой, двоеточия, точки с запятой, тире и точки, при

этом длина предложения не должна предполагать использование более 6-8 знаков препинания в одном предложении.

14. Не допускается использование в пределах одного текста (а по возможности –во всех документах, содержащих правовые предписания или официальную информацию) одного слова в разных значениях или разных слов в одном значении.

15. Слова должны использоваться в общеупотребительном значении, которое будет понятно любому гражданину, не имеющему специальной подготовки и специальных знаний. В случае, если слово, имеющее общеупотребительное значение, используется как специальный термин, значение которого ограничивается или не совпадает с общеупотребительным значением этого слова, это должно специально оговариваться в тексте документа, а специальное терминологическое значение слова должно быть четко обозначено.

16. При использовании близких по смыслу понятий должна быть четко обозначена разница между этими понятиями, например, через дефиниции обозначающих эти понятия терминов.

17. При использовании словосочетания, состоящего более чем из 5 слов, более чем трижды в одном документе его следует заменять кратким термином, соответствие которого обозначаемому им понятию должно приводится либо в глоссарии, либо при первом использовании этого термина. При этом следует избегать использования аббревиатур, кроме общепринятых.

18. Следует избегать использования более 3 однородных членов предложения подряд.

19. Рекомендуется в процессе подготовки проекта любого документа проверять определенность и понятность официальных текстов для их адресатов. Доступность содержания текста для восприятия может оцениваться с помощью подсчета индекса читабельности текста и других инструментов количественного (формального) анализа. Понятность и определенность текста оценивается с помощью проведения анкетирования, опросов или интервью с представителями аудитории, выступающей адресатом данного текста.