Исследование сложности предложений, составляющих тексты правовых актов органов власти Российской Федерации

Д.А. Савельев

научный сотрудник Института проблем правоприменения при Европейском университете в Санкт-Петербурге, кандидат юридических наук. Адрес: 191187, Российская Федерация, Санкт-Петербург, Гагаринская ул., 6/1. E-mail: dsaveliev@eu.spb.ru

Аннотация

Для качественной правореализации недостаточно только факта официального опубликования нормативных актов. Важна ясность правовых текстов, доступность их для понимания. Лингвистическое и юридическое качество текста взаимосвязаны. Создание качественного текста будет способствовать более четкому формулированию идей, заложенных в правовой или судебный акт. В статье содержатся методика и результаты исследования текстов законодательства России, проведенного в целях совершенствования правореализации и правоприменения, снижения затрат времени на восприятие правовых норм, улучшения качества правовых актов. Использованы тексты 199 тыс. правовых актов. Проведена сегментация текстов на 5,5 млн. предложений; автоматизированная морфосинтаксическая разметка предложений с выделением частей речи и их свойств. На этой основе рассчитаны метрики лексической и синтаксической сложности каждого предложения: длина, лексическое разнообразие, длины зависимостей частей речи, длины слов в слогах и др. Выбраны метрики, позволяющие количественно оценить сложность предложений правового текста, которая отличается от литературного текста. Предложена методика автоматизированного определения предложений, которые можно отнести к наиболее трудночитаемым, без использования ручного труда. На основе этой работы созданы и опубликованы примеры плохо читаемых предложений правовых актов. Сведения о предложениях проанализированы статистически. Определены органы власти, которые пишут сложнее, и тематики документов, в которых встречается больше сложно написанных предложений. Показано, что число длинных предложений в законодательстве существенно (в пять раз) возросло по сравнению с первыми годами современной российской государственности. В частности, половина предложений актов Конституционного Суда состоит более чем из 40 токенов каждое. Выделены наиболее часто встречающиеся словосочетания и обороты, которые характеризуют тематику текстов, в которых встречаются наиболее сложные предложения. Опубликованный информационный ресурс может стать в дальнейшем предметом для более детальных работ по совершенствованию юридической техники и содержания правовых и судебных актов.

Ключевые слова

правовая информация, законотворчество, законодательный процесс, правовой акт, корпус, лингвистика, экспертиза, лексическое разнообразие, открытые данные, вычислительная лингвистика, анализ текста.

Благодарности: Публикация подготовлена в рамках научного проекта N 17-18-01618, поддержанного Российским научным фондом.

Автор выражает признательность за методическую помощь Д. Скугаревскому, Р. Кучакову и всем сотрудникам Института проблем правоприменения при Европейском университете в Санкт-Петербурге, давшим рекомендации по осуществлению исследования.

Для цитирования: Савельев Д.А. Исследование сложности предложений, составляющих тексты правовых актов органов власти Российской Федерации // Право. Журнал Высшей школы экономики. 2020. № 1. С. 50–74.

УДК: 340 DOI: 10.17323/2072-8166.2020.1.50.74

Введение

Для обеспечения качественной правореализации недостаточно только лишь факта официального опубликования нормативных актов. Важна ясность правовых текстов, доступность их для понимания. Иначе норма ч. 3 ст. 15 Конституции Российской Федерации об обязательности официального опубликования правовых актов, затрагивающих права человека и гражданина, не имеет смысла. Более того, даже если на время принять крайнюю точку зрения и считать, что тексты правовых документов вовсе не адресованы рядовому гражданину, а являются профессиональным инструментом, предназначенным исключительно для подготовленных юристов, все равно встает вопрос о затратах времени и усилий юристов и руководителей организаций на восприятие и понимание текстов правовых и судебных актов. Улучшение качества текстов и снижение времени на их понимание в целом работают на государство и общество, сокращая ненужные издержки.

Важна связь лингвистического и юридического качеств текста. Создание оптимального с точки зрения лингвистики правового текста также будет способствовать и более четкому формулированию идей, заложенных в правовой акт. Уместно помнить выражение философов прошлого — «кто ясно мыслит, тот ясно излагает». Л. Фуллер писал: «Туманное и бессвязное законодательство может сделать законность недосягаемой для кого бы то ни было или по крайней мере недосягаемой без неправомочного пересмотра, который сам по себе наносит ущерб законности. Воду из загрязненного источника порой можно очистить, но тогда она станет чем-то иным» [Фуллер Л., 2007: 81].

В.Б. Исаков отметил: «Разумеется, лингвистическая экспертиза не всесильна. Никакой редактор не в состоянии превратить бессмысленный набор слов в продуманный и стройный законопроект», и выразил надежду, что в последующем профессионализм законодателя в сфере формулирования правовых текстов возрастет [Исаков В.Б., 2000: 72–89]. Эта надежда пока не оправдалась.

Тема понятности правового акта не раз находила отражение в российской и зарубежной литературе. Движение за простой английский юридический язык возникло в середине прошлого века. Р. Асси утверждает, что жалобы на трудность юридического языка столь же стары, как и право; в последние три десятилетия правительства и корпорации затратили существенные ресурсы на то, чтобы «демистифицировать право для простых людей» при помощи упрощения правовых текстов. Он, тем не менее, отмечает, что упрощение языковых конструкций само по себе не ведет к тому, чтобы «право говорило напрямую с человеком» [Assy R., 2011: 376–404].

В США действует Закон о ясном языке (Plain Writing Act of 2010)[1], который обязывает федеральные органы исполнительной власти формулировать все обращаемые к публике заявления в соответствии со руководством по простому письму. На официальном сайте[2] размещены такое руководство, примеры и инструкции.

Есть много исследований читаемости юридических текстов при помощи лингвистических метрик с точки зрения оценки читаемости асессорами. Например, Де Фриз исследовал в диссертации [De Friez B., 2017] читаемость судебных решений Верховного суда штата Айдахо с 1890 г. по наши дни. Он использовал несколько разных подходов. Самый, пожалуй, распространенный — это индексы удобочитаемости Флеша[3], для которых был применен подсчет коэффициента корреляции Пирсона[4] с годом принятия акта. Эти тесты показали ухудшение показателей читаемости. Однако автор использо-

[1] Available at: https://www.congress.gov/bill/111th-congress/house-bill/946 (дата обращения: 22.10.2019)

[2] Available at: http://plainlanguage.gov (дата обращения: 22.10.2019)

[3] Индекс удобочитаемости Флеша позволяет различать документы по сложности для прочтения. Он рассчитывается по формуле, включающей число слов, число предложений и число слогов в текстах. Индекс Флеша-Кинсайда позволяет ранжировать документы по степени сложности, сравнивая их с уровнями обучения (классы школы, курсы университета и пр.).

[4] Pearson product-moment correlation coefficient используется для определения зависимости изменения одних величин от изменения других (корреляции), в данном случае, года принятия акта и метрик удобочитаемости, рассчитанных по приведенным выше формулам. См., напр.: Pearson Coefficient of Correlation Explained. URL Pearson Coefficient of Correlation Explained. Available at: https://towardsdatascience.com/pearson-coefficient-of-correlation-explained-369991d93404 (дата обращения: 10.02.2020)

вал многофакторный анализ стиля на основе некоторых особенностей языка (например, использования слов в сложных формах и пр.), который дал противоположный результат. Затем был использованы различные модели оценки сложности текста. В итоге автор пришел к заключению, что читаемость текстов улучшилась, хотя они стали длиннее.

Колман и Финг исследовали индексы удобочитаемости 9 тыс. текстов Верховного Суда США с 1969 по 2004 гг., сделав вывод, что читаемость улучшилась за годы, прошедшие с момента возникновения движения за простой язык [Coleman B., Phung Q., 2010: 75]. Оуэнс и Видикин исследовали читаемость документов отдельных судей Верховного Суда США, установив различия между ними, а также приходя, среди прочего, к выводу, что судьи пишут сложнее, когда они пересматривают прецедент [Owens R.J., Wedeking J.P., 2011: 1027–1061].

Исследования юридического языка проводятся и в Европе, и в других регионах. Вальтль и Маттес оценивают читаемость более 3 тыс. законодательных актов Германии при помощи метрик читаемости, а также сравнивают законодательство Германии и Австрии об ответственности за некачественный товар [Waltl B., Matthes F., 2015: 10]. Смит и Ричардсон приводят сведения о большом количестве исследований читаемости австралийских правовых актов в сфере налогообложения [Smith D., Richardson G., 1999: 1027–1061]. Исследование использования юридических жаргонизмов, сложных и устаревших слов привело к выводу, что органам власти объединенной Европы также следует обратиться к использованию более простого языка [Giampieri P., 2016: 424–439]. Также отметим вышедший недавно сборник статей «Право как данные: вычисления, текст и будущее правового анализа» [Livermore M., Rockmore D., 2019: 526], в который вошли 17 материалов, посвященных анализу правовых данных, в том числе стилистики правового текста.

В отечественной юридической литературе много внимания уделяется экспертизе проектов нормативных актов с точки зрения использования языковых конструкций. Государственная Дума опубликовала «Методические рекомендации по лингвистической экспертизе» [Крюкова Е.А., Крыжановская Л.А., 2013: 40], в которых указывается, что приоритет отдается простым предложениям, прямому порядку слов и расположению определений рядом с определяемым словом. Также отмечается, что обратный порядок слов затрудняет чтение. Прямым порядком слов называется формула «подлежащее, сказуемое, второстепенные члены».

Т.В. Губаева [Губаева Т.В., 2004: 38–43] также показывает примеры излишне сложных словесных конструкций и нарушений правил русского языка в нормативных актах, вплоть до законов и кодексов Российской Федерации. Она обращает внимание на то, что некоторые «канцелярские штампы», т.е.

слова и выражения, вроде бы построенные по принципам официально-делового стиля, не несут никакой смысловой нагрузки и лишь загромождают текст правового акта.

В России также создается лингвистически размеченный корпус локальных документов и актов CorRIDA. Авторы ставят целью проекта описание локальных документов разных типов через выделение и сравнение их языковых черт, а также оценку официально-деловых текстов с точки зрения их языковой сложности, удобства для восприятия и понимания «простым носителем» русского языка [Белов С.А. и др., 2018: 114–123].

В.В. Шашек и Н.А. Харченко исследуют при помощи испытуемых понятность норм Гражданского кодекса Российской Федерации (далее — ГК РФ). Испытуемым предлагалось ответить на вопрос «О чем данная статья?» применительно к одной статье ГК РФ (ст. 1150 или 1508). Полученные ответы обобщались, на основе чего был сделан вывод: непрофессиональный читатель не может уяснить смысл такого текста. Затем авторы создали адаптированный текст указанных статей, показав, что возможно понятное изложение норм [Шашек В. В., Харченко Н. А. 2016: 1191–1196].

А.В. Поляков отмечает, что только при ясном изложении правовой нормы она может быть исполнена [Поляков А.В., 2009: 18]. Даже если считать, что развитие законодательства предполагает его конкретизацию [Степанов О.А., 2018: 4–23], то такая конкретизация, на наш взгляд, не должна приводить к ухудшению качества текста.

Опубликование правовых актов в электронной форме на Официальном интернет-портале правовой информации позволило создать информационный ресурс, состоящий из текстов, пригодных для машинного анализа. Институт проблем правоприменения проводит исследование текстов, входящих в него, методами вычислительной лингвистики.

Наши предыдущие исследования показали на уровне всех текстов ухудшение среднегодовых показателей метрик сложности текста для чтения во времени, в особенности, с 2014 года. При этом не подтвердилось предположение, что язык в целом в общественном дискурсе становится сложнее: сравнимая выборка текстов СМИ не показывает ухудшения читаемости текстов. Среди правовых актов обнаружены огромные документы, представляющие собой перечни, например, географических координат или адресов, выгруженные из баз данных, которые в принципе нельзя рассматривать с точки зрения читаемости. Однако это лишь первые результаты, которые носили наиболее общий характер.

В настоящей статье будут показаны результаты подробного исследования текстов правовых актов на уровне предложения для ответа на вопрос, какие именно предложения юридического текста приводят к ухудшению па-

раметров читаемости текста. Также будут приведены примеры таких предложений и сделаны предположения относительно причин этого явления и методов улучшения ситуации.

1. Данные массива правовых текстов и их подготовка

Исходными данными для исследования стал информационный ресурс, состоящий из текстов вновь принятых правовых актов за 1990–2017 гг. Russian Law as Open Data[5], созданный ранее Институтом проблем правоприменения. Этот ресурс сформирован на основе документов, доступных на Официальном интернет-портале правовой информации[6]. В него включены только первоначальные редакции правовых актов на момент их принятия, а также последующие акты об изменениях. Консолидированные версии документов с внесенными изменениями не использовались во избежание двойного подсчета. В настоящем исследовании используется часть указанного ресурса объемом 199 тыс. текстов, которая получена из HTML-текстов раздела портала «Законодательство России».

Важным этапом работы является предварительная обработка текстов для увеличения точности их анализа. В массиве правовых актов встречаются документы огромных объемов. В них может быть как много предложений, так и одно предложение длиной более нескольких тысяч слов. Встречаются документы и предложения, которые в принципе невозможно оценить с точки зрения читаемости как текст: например, Постановлением Правительства России № 534 от 25.07.2013 была утверждена таблица из 1600 страниц, в которой содержатся только координаты точек на карте, обозначающие границы Сочинского национального парка. Такие документы были также исключены из рассмотрения. Документы, содержащие слово «бюджет» в названии, были также полностью исключены из анализа.

Из текстов были удалены таблицы (в них трудно выделить предложения целиком) и части, которые представляют собой формы для заполнения (предложения, в которых оставлены места для заполнения, чаще всего обозначенные большим числом знаков подчеркивания). В документах текст статей и пунктов отделен от служебных частей (удалены начало с названием и сведениями о принятии, окончание с подписью, заголовки и подзаголовки структурных элементов, ссылки на источники опубликования правовых актов и т.п.). Из документов также удалены ссылки на источники опубликования правовых актов.

[5] Available at: URL: https://github.com/irlcode/RusLawOD (дата обращения: 22.10.2019)

[6] Available at: URL: http://www.pravo.gov.ru (дата обращения: 22.10.2019)

Не менее трудной задачей является сегментация правового текста на предложения. В правовых текстах нет предложений, оканчивающихся восклицательным или вопросительным знаком, поэтому использовалось разделение предложений только по наличию точки в тексте. Для качественной сегментации в текстах выделялись сокращения, даты, числа, записанные через точку, и такие точки не считались концом предложения. Списки и перечисления, в которых элементы выделены с новой строки, рассматривались как отдельные предложения.

Для вычисления метрик была проведена морфосинтаксическая разметка предложений. В ходе этой разметки для каждого слова и для каждого знака препинания определен ряд лингвистических свойств (начальная форма слова, использованная форма слова, место по отношению к главному и т.п.)[7]. Для работы использовались морфологический анализатор MyStem, программа частеречной разметки TreeTagger и анализатор зависимостей MaltParser.

2. Характеристика и значение лингвистических метрик для юридического текста

Вычислительная лингвистика представляет возможность автоматизированного определения и подсчета множества метрик, связанных с текстом и предложением. Например, для русского языка Рейнольдс выделяет 179 различных метрик текста. Среди них он при помощи статистического метода выбирает 30 метрик, наиболее ценных по количеству информации для классификации сложности текстов [Reynolds R., 2016: 172].

Задача анализа правовых текстов имеет специфику. Тексты правовых актов существенно отличаются от литературных текстов и разговорного языка. Такие тексты менее разнообразны и существенно формализованы по стилю и содержанию. Если предположить, что правовой акт скорее всего будет читать подготовленный читатель, профессионал, имеющий базовые знания юридической терминологии, то для определения сложности текста можно не учитывать сложности терминологии. Поэтому, не затрагивая на данном этапе сложности используемых терминов, мы остановились на исчислимых лингвистических характеристиках текстов.

Опираясь на приведенное выше статистическое исследование, тем не менее необходимо выбрать именно метрики, которые в наибольшей степени относятся к качеству юридического текста. Это можно сделать, учитывая смысл и значение той или иной метрики для особенностей правового тек-

[7] Данная работа проведена автоматизированным образом с помощью программ ru-syntax (Available at: URL: https://github.com/tiefling-cat/ru-syntax), подготовленного кафедрой компьютерной лингвистики НИУ ВШЭ.

ста. Далее мы приведем подробные сведения о выбранных исходя из этого метриках.

Если разбить предложение на токены, можно получить его длину в токенах. Под токенами понимаются слова, знаки препинания, числа[8]. Длина предложения в целом показывает сложность его восприятия. В правовых актах предложения бывают гораздо длиннее, чем в литературных текстах (например, более 100 слов), и такие предложения даже без дополнительного анализа можно назвать сложными и требующими возможного пересмотра. Отдельно можно выделить токены, которые представляют собой самостоятельные части речи: существительные, прилагательные, глаголы и наречия. Длину предложения можно высчитать в таких словах, опуская знаки препинания, числа и служебные части речи. Однако длины предложения недостаточно, чтобы показать все аспекты сложности предложений даже без учета смысла слов.

Такая характеристика, как лексическое разнообразие, представляет информацию о количестве одинаковых слов в предложении. Можно сказать, что она показывает, сколько раз допущены повторы одних и тех же слов. В юридических текстах часты повторы, которые объясняются строгостью формулировок различных названий субъектов права и составных терминов. Однако перегруженные повторами предложения трудно воспринимать, даже если они построены без ошибок с точки зрения правил русского языка и юридической техники. Поэтому мы выделяем этот вид метрик как один из основных для нашего исследования.

Лексическое разнообразие можно рассчитать различными способами, в том числе считая любые токены (type/token ratio, TTR) или только их определенные виды. При делении общего числа токенов на число уникальных токенов получается значение метрики, изменяющееся от 1 (все слова разные) до стремящейся к 0 величины. В данном исследовании для подсчета метрики предложений было использовано как количество уникальных токенов в целом (TTR), так и отдельно слов, представляющих собой самостоятельные части речи, приведенных к начальной форме слова, которое поделено на общее количество таких слов в предложении (content lemma type/token ratio, CTlemTR). Это в итоге показывает разнообразие слов в предложении независимо от знаков пунктуации, предлогов, союзов и т.п.

В обычных текстах, например, при обучении русскому языку детей определенных возрастных групп или при обучении взрослых другому языку как иностранному увеличение лексического разнообразия показывает увели-

[8] В целом под токеном понимается выделенный в тексте минимальный фрагмент для последующего анализа (слово, число, знак препинания и т.д.). Здесь для получения токенов тексты разбиваются через пробел и отделяются знаки препинания.

чение сложности текста. Чем больше разных новых слов встречается читающему, тем сложнее ему воспринимать текст. Однако в правовых текстах частые повторы слов (меньшее разнообразие) дают обратный эффект — затруднение читаемости. Поэтому указанная метрика используется для изучения ухудшения читаемости за счет повторов слов в тексте.

Значение этой метрики связано с длиной предложения: чем длиннее предложение, тем больше вероятность повтора слов. В текстах для уменьшения влияния этого эффекта применяется сэмплирование (выбор отрезков текста одинаковой длины). Однако такой способ в случае предложений не имеет смысла. Тем не менее мы считаем измерение этой метрики допустимым для определения качества текста. Данная метрика покажет не столько чистое лексическое разнообразие, сколько лексическое разнообразие в его связи с длиной предложения. Длина предложения в правовом тексте также сказывается на его читаемости.

Максимальное расстояние между связанными частями предложения (maxDepLen). Существует группа метрик, которые характеризуют сложность построения структуры предложения независимо от смысла каждого слова, входящего в него, или от их разнообразия. Структура предложения может быть изображена в виде дерева, в котором установлена связь подчинения между зависимыми членами предложения. О том, что метрики, связанные с деревом зависимостей, можно использовать для анализа юридического текста, уже давно упоминалось в литературе [Костенко М.А., 2005: 127]. Однако сейчас благодаря компьютерной лингвистике и машинному обучению стало возможно строить такие деревья не ручным способом по каждому предложению, а массово, на большом количестве текстов, при помощи автоматической обработки текста. В дискретной математике разработана и с успехом используется теория графов. Под графом понимается структура, состоящая из точек (вершины) и соединяющих их связей (рёбра). Такая структура может описать множество различных сетей: например, сети связи, сети железных дорог, и т.д. Она может использоваться для работы с деревом зависимостей членов предложения в тексте. Для работы с анализом графов разработано множество методов и соответствующее программное обеспечение, которое может использоваться и в нашем случае[9]. В рамках исследования дерева структуры предложения можно выделить несколько различных параметров, характеризующих это дерево. Так, можно исследовать длину и количество его ветвей в общей сумме, вычислять среднее или максимальное значение такой длины. Наиболее интересной метрикой, дающей больше информации о сложности текста, является максимальное число слов, лежащих

[9] В настоящем исследовании применена библиотека NetworkX для языка программирования Python.

между зависимыми членами предложения (maximum Dependency Length). Эта метрика говорит о том, сколько слов читатель должен «перескочить» в предложении, чтобы уяснить его смысл.

Предложение становится трудным для восприятия, если оно перегружено длинными словами. Ранжировать слова по сложности при этом можно, рассчитывая длину слова в слогах[10]. От использования некоторых длинных слов-терминов отказаться невозможно, но число длинных слов в предложении не должно быть большим. В связи с этим сложность текста также может характеризоваться средним числом слогов на слово. Этот параметр широко используется в формулах удобочитаемости.

Длинные предложения трудно воспринимать. Однако можно различать «технические» предложения, которые содержат только перечисления (списки) различных сущностей (наименований организаций, географических наименований, ФИО лиц, и пр.), и действительно сложно составленные длинные предложения, которые не образуют список.

«Технические» предложения невозможно оценивать с точки зрения их конструкции, а в целом их было бы правильнее оформлять не как предложение текста, а как список с нумерацией, таблицу или приложение. Поэтому такие предложения следует выделять.

Но помимо таких «технических» предложений есть и действительно сложно составленные длинные предложения. Авторы текстов могут соединять фактически разные предложения в одно при помощизнаков препинания (запятых, точек с запятой) и соединительных союзов. Это существенно затрудняет читаемость.

Чтобы разделить такие предложения, можно использовать метрику числа соединений, которая означает сумму числа запятых и соединительных союзов. При этом ее соотношение с длиной предложения в словах может выделить списки среди длинных перечислений.

3. Методика выбора трудно читаемых предложений

Каким образом выделить из всего массива предложений такие, которые можно отнести к плохо читаемым на основании приведенных выше метрик? Необходимо определить диапазоны значений метрик, которые характеризуют предложение как плохо читаемое, а также решить вопрос, как сочетать значения разных метрик для отбора предложений.

Один из возможных способов определения — оценка предложений асессорами, т.е. людьми, которые читают случайно выбранные предложения и оценивают их согласно своему опыту. Однако такая методика имеет ряд не-

[10] Предполагается, что количество слогов равно количеству гласных букв.

достатков, например, трудоемкость и субъективизм оценщиков. Кроме того, избранный дизайн не включает оценку смысла слов предложений, а асессорам будет сложно абстрагироваться от сложности смысла слов. Поэтому был избран другой способ — сравнить предложения между собой на основе статистики и отобрать такие, которые имеют худшие значения разных метрик относительно большинства. При такой методике мы опираемся на опыт всех лиц, которые составляли все тексты, и предполагаем, что большинство предложений написаны в среднем достаточно хорошо, а изучения требуют те, которые существенно отличаются от них. Также эта методика позволяет нам сделать выводы, не изучая смысла самих метрик и не прочитывая предложения. Однако сделать это путем сравнения только со средним значением нельзя: необходимо сделать поправку на возможную неоднородность распределения значений, а также заранее нужно убрать из рассмотрения длинные списки.

Поскольку в настоящем исследовании используются несколько разных метрик, каким образом можно определить плохое качество составления предложений правовых актов на основании каждой из них? В настоящем исследовании было выбрано два различных подхода. Во-первых, более широкая выборка предложений сформирована по одной метрике — длине предложения, однако с фильтрацией длинных перечислений. Во-вторых, еще точнее это возможно сделать, определив пересечение множеств и получить из «широкого» «узкий» массив предложений, обладающих существенной сложностью по всем избранным показателям одновременно. В итоге исследования были сформированы два таких массива предложений.

Избранная на втором этапе методика определения пересечения множеств предложений состоит, во-первых, в автоматической разбивке всех предложений на десять групп по каждой из сравниваемых метрик таким образом, что количество предложений в каждой группе примерно равно[11]. Затем последовательно выявляется число предложений, находящихся одновременно в группе с одним номером и выше по всем трем метрикам. На основании этого выбраны диапазоны значений каждой метрики для предложений, которые могут считаться плохо читаемыми. Эти значения указаны в следующем разделе настоящей статьи.

4. Результаты создания выборки трудно читаемых предложений законодательства

Сегментация всех текстов доступных правовых актов позволила нам выделить примерно 5,5 млн. предложений. При этом практически половина

[11] Эта операция производится функцией Pandas qcut на языке Python.

текстов состоит менее чем из 5 предложений. Средняя длина предложения– примерно 21 токен. Если длину предложения измерить в словах, представляющих собой самостоятельные части речи[12] — примерно 13 слов. Среднее число слогов по всем текстам 2.8. Что касается лексического разнообразия (TTR), более 470 тыс. показывают этот параметр ниже 0.5, что означает, что в них половина слов повторяется, или несколько слов встречаются очень часто.

При том, что средние величины представляются нам небольшими, наблюдаются выбросы по значениям. Примерно 33 тыс. предложений достигают в длину 100 токенов и более. Длина зависимых связей более 30 слов (что существенно затрудняет восприятие) встречается в 285 тыс. предложений. Если предложение состоит из слов, среднее число слогов у которых выше 4, предложение можно считать более трудным для восприятия, чем подавляющее большинство предложений в тексте. Такое предложение состоит преимущественно из длинных слов, например, «совершенствования функционирования комплексных систем обеспечения безопасности жизнедеятельности населения». Подобных предложений 273 тыс., 190 тыс. предложений имеют длину более 40 слов. Для сравнения: в произведении Л.Н. Толстого «Анна Каренина», по нашим подсчетам, средняя длина предложения около 14 слов. Приведенные выше количества предложений невелики (до 5% общего числа), однако законодательные акты и акты судов высших инстанций составляют меньшую часть всех документов. Например, длинные предложения более 40 токенов уже составляют 17% предложений всех федеральных законов и более 51% предложений актов Конституционного Суда Российской Федерации.

В более широкую выборку включено около 40 тыс. предложений длиной более 60 слов, представляющих самостоятельные части речи. Количество 60 слов было выбрано на основании изучения гистограммы распределения количества слов по предложениям (рис. 1).

Выборка этих предложений также ограничена по лексическому разнообразию (разнообразие самостоятельных частей речи находятся в промежутке 0,95–0,4 и общее лексическое разнообразие (TTR) находятся в промежутке 0,95–0,29). Такие ограничения введены с учетом необходимости отбросить длинные списки различных сущностей. Эта выборка может быть самостоятельным предметом для изучения с точки зрения проверки юридической техники, смысла предложения, уместности длинных перечислений в одной строке.

Как мы видим, предложения длиной более 60 слов — самостоятельных частей речи — встречаются очень редко по сравнению с общей массой, в связи с чем было выбрано такое значение.

[12] В данном случае это существительные, прилагательные, глаголы и наречия.

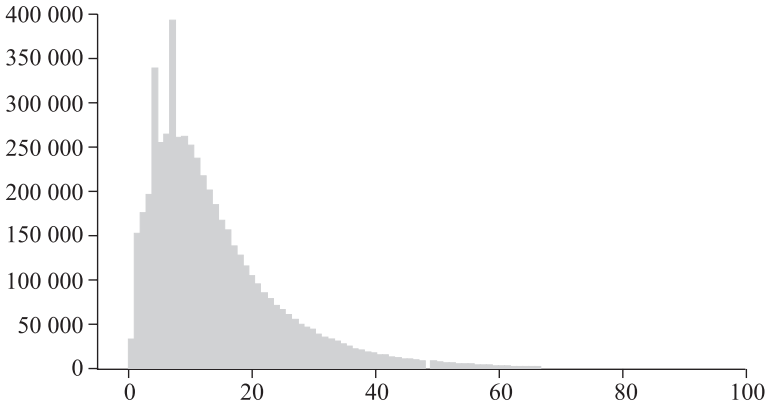

Рис. 1. Гистограмма распределения числа предложений по интервалам значений числа самостоятельных частей речи

Примечание. На рисунке по оси X значения числа слов в предложении, представляющих самостоятельные части речи, разделены на 100 интервалов. Граница интервала по оси X ограничена максимальным значением 100. Реальное значение максимума на таком графике без ограничения составило бы более 3000. По оси Y показано количество предложений, попадающих в соответствующий интервал с учетом указанных выше ограничений выборки.

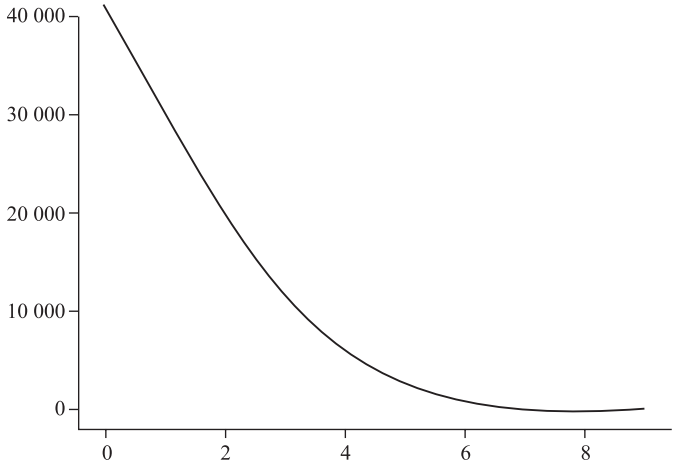

Рис. 2. Число предложений, попадающих в соответствующие десятичные интервалы по трем метрикам

Примечание. По оси X указаны интервалы метрик, разбитые на 10 (децили), а по оси X — количество предложений, попадающих одновременно в указанный дециль и хуже по трем метрикам. Так, в 0 дециль попадают все предложения, в последний (9) — 0 предложений. Численные значения: 40967, 29443, 20091, 11857, 6123, 2454, 782, 187, 18, 0. На основании графика было выбрано значение 6 дециля (787 предложений), однако выбор может быть расширен или сужен для нахождения большего или меньшего числа предложений.

На втором этапе из широкой выборки была выделена более узкая выборка в 782 предложения, которые одновременно входят в число наиболее некачественных по трем измеренным метрикам, показанным выше: максимальная длина зависимостей больше 61; средняя длина слова в слогах больше 2,9; разнообразие самостоятельных частей речи меньше 0,64. Эти промежутки выбраны исходя из анализа распределения количества предложений, попадающих в тот или иной интервал метрик во всей совокупности обработанных предложений, по методике, изложенной в предыдущем разделе.

В итоге сформированы две выборки трудно читаемых предложений. Они опубликованы на свободно доступном ресурсе GitHub13.

5. Примеры текстов и анализ содержания предложений, отнесенных к плохо читаемым

Во избежание существенного увеличения объема статьи мы не можем привести здесь примеры наиболее длинных предложений, длина которых исчисляется сотнями слов. Они опубликованы по указанной ссылке. Ограничимся лишь ссылками на документы. Так, например, Постановление Правительства Российской Федерации14 содержит предложение длиной 294 токена. Федеральный закон от 31.12.2007 N 504-ФЗ[15] содержит предложение длиной 298 токенов. В основополагающем в своей отрасли Федеральном законе «Об образовании в Российской Федерации»[16] допущено предложение с длиной дерева зависимостей более 200 слов и разнообразием менее 0,1 и т.д.

Два не столь длинных предложения (около 65 токенов в длину) можно привести в пример для наглядной иллюстрации низкого лексического разнообразия за счет повторов слов. Из федерального закона:

«В целях координации деятельности и контроля за выполнением соглашения о создании **территории опережающего социально-экономического развития**, содействия в реализации проектов резидентов **территории опе-**

---

[13] Как подраздел информационного ресурса Russian Law as Open Data. Available at: URL: https://github.com/irlcode/RusLawOD (дата обращения: 23.10.2019). Обе указанные части представляют собой таблицы, в строке которых приведено предложение и все измеренные метрики по этому предложению, а также метаданные (из какого документа извлечено предложение).

[14] Постановление «Об утверждении Правил предоставления (использования, возврата) из федерального бюджета бюджетам субъектов Российской Федерации бюджетных кредитов на 2016 год» от 27.01.2016 N 40 // СЗ РФ. 2016. N 5. Ст. 708.

[15] Федеральный закон «О внесении изменений в Федеральный закон «О контрактной системе в сфере закупок товаров, работ, услуг для обеспечения государственных и муниципальных нужд» от 31.12.2007 N 504-ФЗ // СЗ РФ. 2018. N 1 (ч. I). Ст. 88.

[16] Федеральный закон «Об образовании в Российской Федерации» от 29.12. 2012 N 273-ФЗ // СЗ РФ. 2012. N 53 (ч. I). Ст. 7598.

режающего социально-экономического развития, проектов иных инвесторов, оценки эффективности функционирования **территории опережающего социально-экономического развития**, а также в целях рассмотрения и утверждения перспективных планов **развития территории опережающего социально-экономического развития**, осуществления контроля за реализацией этих планов создается наблюдательный совет **территории опережающего социально-экономического развития**»[17].

Из подзаконного нормативного акта: «Решение о выплате единовременного поощрения **руководителю** территориального органа Росаккредитации принимается **Руководителем**, заместителю **Руководителя** принимается **Руководителем**, решение о выплате единовременного поощрения **руководителю** структурного подразделения принимается **Руководителем** по представлению заместителя **Руководителя**, осуществляющего координацию деятельности структурного подразделения, решение о выплате единовременного поощрения другим гражданским служащим Росаккредитации принимается Руководителем по представлению **руководителя** структурного подразделения, согласованного с заместителем **Руководителя**, осуществляющим координацию деятельности структурного подразделения»[18].

Как мы видим, повторы одних и тех же слов едва ли позволяют после первого прочтения уяснить смысл предложения, хотя сам смысл не сложен и предложение не содержит ошибок в использовании языка. Рассмотрев эти предложения, отметим два аспекта проблемы: во-первых, не изменяя ничего в существе их текста, технически текст можно было бы переписать гораздо проще и понятнее; во-вторых, в первом предложении допущены абстракции — координация, контроль, попытка дать создаваемому органу полномочия, но не вполне определенная. Это касается уже не только технического, но и юридического качества текста. Важно подчеркнуть, что недостатки технического качества предложений сопутствуют некачественным в правовом смысле текстам.

Время, необходимое для прочтения и осознания сложных предложений, является критическим фактором, в связи с которым не следует допускать

---

[17] Федеральный закон «О территориях опережающего социально-экономического развития в Российской Федерации» от 29.12.2014 N 473-ФЗ // СЗ РФ. 2015. N 1 (ч. I). Ст. 26.

[18] Из Приказа Росаккредитации от 05.02.2018 N 22 «Об утверждении Положения о порядке выплаты ежемесячной надбавки к должностному окладу за особые условия федеральной государственной гражданской службы, ежемесячной надбавки к должностному окладу за выслугу лет на федеральной государственной гражданской службе, материальной помощи, единовременной выплаты при предоставлении ежегодного оплачиваемого отпуска, премирования за выполнение особо важных и сложных заданий, единовременного поощрения за безупречную и эффективную федеральную государственную гражданскую службу федеральным государственным гражданским служащим Федеральной службы по аккредитации» . Available at: URL: http/:www.pravo.gov.ru, N 0001201803210011 (дата обращения: 21.03.2018)

сложных предложений. Автор не проводил масштабной оценки времени понимания найденных предложений асессорами, однако для примера было взято одно из предложений и измерено время чтения — более минуты. В ходе одного прочтения его содержание все равно было невозможно установить: "… предписания, обязательные для исполнения территориальной сетевой организацией, оказывающей услуги по передаче электрической энергии, организацией, осуществляющей холодное водоснабжение и (или) водоотведение (организация водопроводно-канализационного хозяйства), организацией, осуществляющей горячее водоснабжение, газораспределительной организацией, теплоснабжающей организацией при осуществлении деятельности в рамках исчерпывающих перечней процедур в сферах строительства, утвержденных Правительством Российской Федерации в соответствии с частью 2 статьи 6 Градостроительного кодекса Российской Федерации, о совершении действий, направленных на устранение нарушений порядка осуществления в отношении юридических лиц и индивидуальных предпринимателей, являющихся субъектами градостроительных отношений, процедур, включенных в исчерпывающие перечни процедур в сферах строительства, в том числе предписания о заключении договоров, об изменении условий договоров или о расторжении договоров в случае, если лицами, права которых нарушены или могут быть нарушены, было заявлено соответствующее требование…"[19].

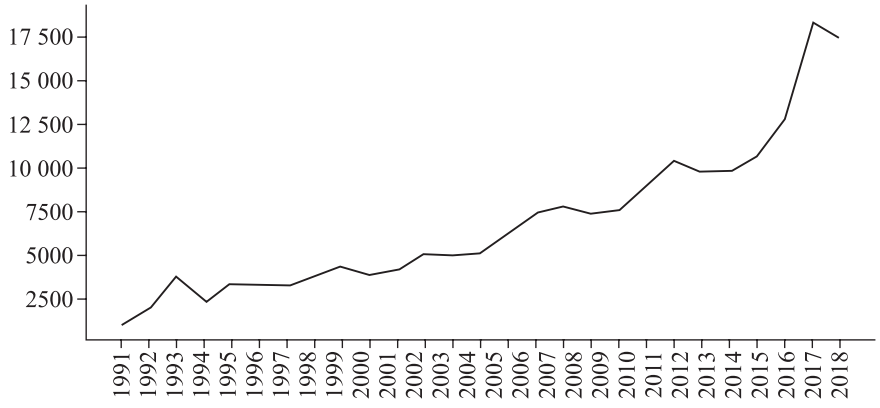

Рис 3. Количество предложений, состоящих более чем из 40 слов по годам

Примечание. По оси X указан год подписания правового акта, по оси Y указано среднее число предложений, состоящих из более чем 40 слов, представляющих самостоятельные части речи.

[19] Из Постановления Правительства РФ от 25.12.2015 N 1429 «Об изменении и признании утратившими силу некоторых актов Правительства Российской Федерации» // СЗ РФ. 2016. N 1 (ч. II). Ст. 239.

Выборку плохо читаемых предложений также проанализируем статистическими методами.

Сравнив число длинных предложений в законодательстве за каждый год, мы видим, что оно существенно (в пять раз) возросло по сравнению с первыми годами современной российской государственности (рис. 1). Количество ежегодно принимаемых актов увеличивается объективно, но если бы предложения были составлены качественно, оно не росло бы из года в год.

Предполагая, что слишком длинные предложения ухудшают качество текста для восприятия, можно связать данные об органе власти, принимающем правовые акты, с массивом предложений (по идентификатору документа, в котором обнаружено предложение). Так можно увидеть органы власти, принимающие плохо читаемые предложения (табл. 1.

Таблица 1

Виды документов по наибольшему числу предложений длиной более 40 токенов

Вид документа	Предложений всего	Предложений больше 40 токенов
Постановление Конституционного Суда	26181	28938 (53%)
Определение Конституционного Суда	21922	22008 (50%)
Приказ Федерального казначейства	9360	2126 (19%)
Приказ Федеральной службы по финансовым рынкам	26108	5551 (18%)
Федеральный закон	390926	83011 (18%)
Постановление Совета Федерации	73611	14893 (17%)
Федеральный конституционный закон	6628	1278 (16%)
Постановление Совета Министров РСФСР	8175	1567 (16%)
Приказ Министерства по налогам и	18402	3502 (16%)
Приказ Федеральной службы по тарифам	7214	1339 (16%)
Приказ Следственного комитета	5519	1022 (16%)
Приказ Министерства финансов	70328	12992 (16%)
Приказ Министерства по антимонопольной политике и поддержке предпринимательства	5959	1068 (15%)
Приказ Министерства строительства и жилищно-коммунального хозяйства	8815	1547 (15%)
Приказ Государственной фельдъегерской службы	7867	1355 (15%)
Приказ Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека	6435	1057 (14%)
…		

Вид документа	Предложений всего	Предложений больше 40 токенов
Постановление Правительства	924297	146726 (14%)
Кодекс	40566	5484 (12%)
Закон Российской Федерации*	17539	2034 (10%)
…		
Распоряжение Президента	55928	3723 (6%)
Указ Президента	644981	23639 (4%)

Примечание. Таблица приведена с сокращениями. Закон Российской Федерации — вид законодательного акта, принимавшийся до вступления в силу действующей Конституции России, после чего стали приниматься федеральные законы.

Подсчет количества предложений по тематикам Общеправового классификатора законодательства, связанным с соответствующими документами, дает такое распределение по тематикам (табл. 2).

Таблица 2

Количество трудночитаемых предложений по тематикам

Код тематики	Название тематики	Количество
010.140.030.010.000	Отмена, изменение и дополнение нормативных правовых актов	11901
	Не определено	10345
010.140.040.025.000	Соответствие Конституции Российской Федерации федерального законодательства	5800
010.140.040.045.040	Правила, инструкции, указания, порядки и иные решения	4898
020.050.000.000.000	Обращения, заявления и жалобы граждан	3593
080.080.030.020.000	Формы финансовой помощи (субвенции, дотации, ссуды, субсидии)	2218
030.030.040.020.060	Акционерное общество	1655
010.140.040.045.020	Иные положения (Учет и систематизация нормативных правовых актов)	1641
020.010.020.020.000	Компетенция (Правительство Российской Федерации)	1631
080.050.010.000.000	Общие положения (Федеральный бюджет)	1431
090.010.070.020.000	Электроэнергия. Распределение, лимиты	1269
120.030.080.000.000	Предоставление информации. Информационные услуги	1244

Код тематики	Название тематики	Количество
020.030.020.040.000	Целевые программы (комплексные)	1099
020.010.040.030.250	Министерство финансов Российской Федерации	1083
080.100.010.000.000	Общие положения (Налоги и сборы)	1075
010.150.010.000.000	Общие положения (Местное самоуправление)	1041
080.060.020.000.000	Доходы бюджетов субъектов Российской Федерации	1023
080.110.020.010.000	Центральный банк Российской Федерации	994
080.080.020.020.000	Пенсионный фонд Российской Федерации	979
180.060.010.000.000	Общие положения (Уголовный процесс)	945

Примечание. В таблице представлены тематики Общеправового классификатора законодательства в той форме, как они опубликованы на Официальном интернет-портале правовой информации. У одного документа может быть более одной тематики, и в одном документе может быть более одного предложения, попавшего в выборку.

Современный метод извлечения информации из текстов[20] позволяет выделить наиболее устойчивые словосочетания, характерные для всего массива предложений. Это делает возможным определить, какие словосочетания в нем повторяются чаще всего (табл. 3).

Таблица 3

Словосочетания, наиболее часто встречающиеся в плохо читаемых предложениях

Словосочетание	Число
Российской Федерации	10957
том числе	2682
федерального бюджета	674
установленном порядке	663
Федеральным законом	536
Постановлением Правительства Российской Федерации	513
местного самоуправления	498
субъектов Российской	487
электрической энергии	476

[20] Для решения этой задачи использовано программное обеспечение Gensim (Available at: https://radimrehurek.com/gensim/models/phrases.html (дата обращения: 10.02.2020)), основанное на работах [Mikolov et al., 2013: 3111-3119]; [Bouma, G, 2009:31-40].

Словосочетание	Число
федерального закона	470
может быть	464
субъекта Российской	432
Конституционный Суд Российской Федерации	423
субъектов Российской Федерации	395
соответствии законодательством Российской Федерации	394
юридического лица	377
таким образом	375
заменить словами	372
Правительства Российской	367
ценных бумаг	357
федерации далее	353
Конституции Российской Федерации	348
федеральных органов исполнительной власти	344
субъекта Российской Федерации	339
юридических лиц	328
территории Российской Федерации	326
акционерного общества	322
МВД России	319
внутренних дел Российской Федерации	312
денежных средств	306
ликвидации последствий стихийных бедствий	306
государственной регистрации	306
муниципального образования	304
целях обеспечения	302

Примечание. В таблице показано количество найденных коллокаций (словосочетаний), наиболее часто употребляемых в выборке трудно читающихся предложений. Для их выделения использован алгоритм взвешивания NPMI, реализованный в программном пакете Gensim. Для выявления коллокаций использовалась предварительная фильтрация стоп-слов и не использовалась лемматизация (приведение к начальной форме слова). Показаны только коллокации, которые встречаются 300 и более раз.

Из данных таблицы следует, что прежде всего и главным образом допускаются повторы словосочетания «Российская Федерация». Также используемый метод дает результаты, похожие на результаты анализа назначенных вручную тематик классификатора, что говорит о высокой эффективности

современных методов извлечения информации из текстов. Использованное программное обеспечение и выбранные параметры также могут варьироваться для решения более узких задач.

Заключение

Для совершенствования правового регулирования, правоприменения и правореализации важно понимать причины, которые ведут ко все большему усложнению предложений в законодательных и судебных актах, несмотря на наличие лингвистической экспертизы, методических рекомендаций. Они имеют не только технико-лингвистические причины. Один из подходов к объяснению виден в анализе текстов решений Конституционного Суда, выполненном А. Дмитриевой, где автор показывает, что ухудшение метрик читаемости происходит в текстах, где отказано заявителю [Дмитриева А.В., 2017: 125–133]. Выше также приводился пример исследования в США, где показана зависимость сложности текста от того, изменяется ли прецедент судом [Owens R., Wedeking J., 2011: 1027–1061]. Таким образом, стилистика юридического текста зависит от его содержания или от особенностей автора. Важную теоретическую базу для этого дает такая наука, как психолингвистика. Например, психолингвистические исследования связывают лексическое разнообразие в речи и искажение смысла сказанного [Dulaney E., 1982: 75–82]. «Формальные» лингвистические метрики могут косвенно говорить о тех или иных мотивах автора текста, таких как создание двойных смыслов текста, сокрытие истинных намерений, или излишнее желание показать свой авторитет.

Исследуя стиль шведского законодателя, Б. Гуннарссон отмечает, что условия, в которых был написан текст, оказывают большое влияние на то, как написан этот текст. Недостаточно проводить лингвистическую экспертизу конечного результата — для того, чтобы создавать более понятные правовые тексты, необходимо реформировать сам процесс их принятия [Gunnarsson B., 1989: 86–107].

Являются ли необходимыми атрибутами официально-делового стиля законодательного текста и текста высших судебных инстанций частые формальные повторы одинаковых словосочетаний, например, «Российская Федерация», и длина выражений, исчисляемая сотнями слов? Или они скрывают мотивы, не отражаемые в тексте, а иногда и некомпетентность составителя текста? Опубликованный информационный ресурс, содержащий трудночитаемые тексты законодательства, может помочь юристам и лингвистам найти причины ухудшения качества текстов и исправить ситуацию.

Библиография

Белов С.А., Блинова О.В., Гулида В.Б. и др. Корпус русских локальных документов и актов CorRIDA: цели формирования, состав, структура / Компьютерная лингвистика и вычислительные онтологии. Сборник статей. Выпуск 2. СПб.: ИТМО, 2018. С. 114–123.

Губаева Т.В. Язык и право. Искусство владения словом в профессиональной юридической деятельности. М.: Норма, 2004. 160 с.

Дмитриева А.В. «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации // Сравнительное конституционное обозрение. 2017, N 3. С. 125–133.

Исаков В.Б. Язык права / Юрислингвистика-2: Русский язык в естественном и юридическом бытии: межвуз. сб. Барнаул: Алтайский государственный университет, 2000. С. 72–89.

Костенко М.А. Правовая лингвистика в законотворческом процессе // Известия ЮФУ. Технические науки. 2005. N 9. URL: https://cyberleninka.ru/article/n/pravovaya-lingvistika-v-zakonotvorchestvom-protsesse (дата обращения: 22-10-2019)

Крюкова Е.А., Крыжановская Л.А. Методические рекомендации по лингвистической экспертизе законопроектов. М.: Государственная Дума, 2013. 40 с.

Поляков А.В. Язык нормотворчества и вопросы юридической техники / Комментарий к Федеральному закону «О государственном языке Российской Федерации». Часть 1. Доктринальный и нормативно-правовой комментарий. СПб.: СпбГУ, 2009. С. 16–28.

Степанов О.А. О проблеме конкретизации права в условиях цифровизации общественной практики // Право. Журнал Высшей школы экономики. 2018. N 3. С. 4–23.

Фуллер Л. Мораль права. М.: ИРИСЭН, 2007. 308 с.

Шашек В.В., Харченко Н.А. Проблема ясности языка законодательства и множественности интерпретаций текстов законов (на материале статей Гражданского Кодекса Российской Федерации) // Молодой ученый. 2016. N 7. С. 1191–1196.

Assy R. Can the Law Speak Directly to Its Subjects? The Limitation of Plain Language. Journal of Law and Society, 2011, no 3, p. 376–404.

Bouma G. Normalized (pointwise) mutual information in collocation extraction / Proceedings of the Biennial GSCL Conference. Potsdam, 2009, pp. 31–40.

Coleman B., Phung Q. The Language of Supreme Court Briefs: A Large-Scale Quantitative Investigation. J. App. Prac. & Process, 2011, vol. 11, pp. 75–103.

De Friez B. Toward a Clearer Democracy: The Readability of Idaho Supreme Court Opinions as a Measure of the Court's Democratic Legitimacy: PhD thesis. Moscow City, 2017, 144 p.

Dulaney E. Changes in language behavior as a function of veracity. Human Communication Research, 1982, no 1, pp. 75–82.

Engberg J. Legal linguistics as a mutual arena for cooperation: Recent developments in the field of applied linguistics and law. AILA Review, 2013, no 1, pp. 24–41.

Giampieri P. Is the European Legal English Legalese-Free. Italian J. Pub. L., 2016, vol. 8, pp. 424–450.

Gunnarsson B. Text comprehensibility and the writing process: The case of laws and lawmaking. Written communication, 1989, no 1, pp. 86–107.

Lundeberg M. Metacognitive Aspects of Reading Comprehension: Studying Understanding in Legal Case Analysis. Reading Research Quarterly, 1987, no 4, pp. 407–432.

Livermore M., Rockmore D. (eds.) Law as Data: Computation, Text, and the Future of Legal Analysis. Santa Fe: New Mexico Institute Press, 2012, 526 p.

Mikolov T. et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 2013, vol. 26, p. 3111–3119.

Owens R., Wedeking J. Justices and legal clarity: Analyzing the complexity of US supreme court opinions. Law & Society Review, 2011, no 4, pp. 1027–1061.

Reynolds R. Russian Natural Language Processing and Computer-assisted Language Learning. PhD thesis. Tromsø: Universitet, 2016, 172 p.

Smith D., Richardson G. The Readability of Australia's Taxation Laws and Supplementary Materials: An Empirical Investigation. Fiscal Studies, 1999, no 3, pp. 321–349.

Waltl B., Matthes F. Comparison of Law Texts. An Analysis of German and Austrian Legislation regarding Linguistic and Structural Metrics. Paper presented at Internationales Rechtsinformatik Symposium. 2015. Available at: https://wwwmatthes.in.tum.de/pages/1occngdfehma2/Comparison-of-Law-Texts-An-Analysis-of-German-and-Austrian-Legislation-regarding-Linguistic-and-Structural-Metrics (дата обращения: 22-10-2019)

Pravo. Zhurnal Vysshey Shkoly Ekonomiki. 2020. No 1

A Study in Complexity of Sentences Constituting Russian Federation Legal Acts

Denis Saveliev

Researcher, Institute for Implementing Law, European University in Saint Petersburg, Candidate of Juridical Sciences. Address: 6/1 Gagarinskaya Str., Saint Petersburg 191887, Russian Federation. E-mail: dsaveliev@eu.spb.ru

Abstract

To ensure proper law enforcement, the fact of official publication of regulatory acts it is not enough. What is important is the clarity of legal texts, their accessibility for understanding. Linguistic and legal quality of the text are interconnected. Creation of a text that is good from the point of view of linguistics will contribute to a clearer formulation of ideas embodied in a legal or judicial act. Linguistic aid after the creation of the draft act is insufficient. It is necessary to take into account recommendations for the clear writing of texts at the stage of creating a legal act. The methodology and results of a study of Russian legislation texts carried out in order to improve law enforcement and mobilization, to reduce the time spent on the perception of legal norms, and to improve the quality of legal acts are presented. A corpus of texts from 199 thousand legal acts was used. Its texts were segmented into 5.5 million sentences. Using artificial intelligence technologies, morphosyntactic markup of sentences with the allocation of parts of speech and their properties was carried out. On this basis, the metrics of the lexical and syntactic complexity of each sentence were calculated: length, lexical diversity, lengths of dependencies of parts of speech (Dependency Length), word lengths in syllables, etc. Metrics were selected that quantified the complexity of sentences in a legal text, which is different from the literary text. A technique is proposed for the automated search of

sentences that can be attributed to the most difficult to read without the use of manual labor. On the basis of this work, a body of poorly readable sentences of legal acts was created and published in the public domain, consisting of a wider selection — too long sentences and narrower — sentences that differ for the worse from the majority in three metrics at the same time. This corpus is analyzed statistically and the authorities that write more difficult are identified, and the subjects of documents in which there are more complex written sentences. It is shown that the number of long sentences in the legislation has significantly (5 times) increased in comparison with the first years of modern Russian statehood. Half of the sentences from acts of the Constitutional Court of the Russian Federation consist of more than 40 tokens. Using the NPMI method, the most frequently occurring phrases and phrases that characterize the subject of the text are selected from the body. The published corpus may become a subject for more detailed work on improving the legal technique and content of legal and judicial acts.

Keywords

legal information; lawmaking; lawmaking procedure; legal act; corpus linguistics; proofreading; lexical variability; open data; computational linguistics; text mining.

Acknowledgments: The work is supported by the Russian Science Foundation, project N 17-18-01618.

Author thanks for methodological aid to D. Skugarevsky, R. Kuchakov and to all researchers of the Institute for Implementing Law of European University in Saint Petersburg for their recommendations.

For citation: Saveliev D.V. (2020) A Study in Complexity of Sentences Constituting Russian Federation Legal Acts. *Pravo. Zhurnal Vysshey shkoly ekonomiki*, no 1, pp. 50–74 (in Russian)

DOI: 10.17323/2072-8166.2020.1.50.74

References

Assy R. (2011) Can the Law Speak Directly to Its Subjects? The Limitation of Plain Language. *Journal of Law and Society*, no 3, pp. 376–404.

Belov S.A. et al. (2018) *Corpus of Russian local documents and acts CorRIDA: aims, contents, structure.* Saint Petersburg: ITMO Press, pp. 114–123 (in Russian)

Bouma G. (2009) Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference*, pp. 31–40 (in Russian)

Coleman B., Phung Q. (2010) The Language of Supreme Court Briefs: A Large-Scale Quantitative Investigation. J. *App. Prac. & Process*. Vol. 11. P. 75–103.

De Friez B. (2017) *Toward a Clearer Democracy: The Readability of Idaho Supreme Court Opinions as a Measure of the Court's Democratic Legitimacy.* PhD Thesis. Moscow City (Idaho), 144 p.

Dmitrieva A.V. (2017) Art of legal writing: quntitative analysis of decisions of the Russian Constitutional Court. *Sravnitel`noe konstitucionnoe obozrenie*, no 3, pp. 125–133 (in Russian)

Dulaney E. (1982) Changes in language behavior as a function of veracity. *Human Communication Research*, no. 1, pp. 75–82.

Engberg J. (2013) Legal linguistics as a mutual arena for cooperation: Recent developments in the field of applied linguistics and law. *AILA Review*, no 1, pp. 24–41.

Fuller L. (2007) *Moral of law.* Moscow: IRISEN, 308 p. (in Russian)

Giampieri P. (2016) Is the European Legal English Legalese-Free. *The Italian Journal of Public Law,* no 8, p. 424.

Gubaeva T.V. (2004) *Language and law. Art of words in professional legal activity.* Moscow: Norma, 160 p. (in Russian)

Gunnarsson B. (1989) Text comprehensibility and the writing process: The case of laws and lawmaking. *Written communication*, no 1, pp. 86–107.

Isakov V. B. (2000) Language of law. *Yurislingvistika: Russian language in its natural and juridical being.* Barnaul: University, pp. 72–89 (in Russian)

Kostenko M.A. (2005) Legal language in legislative procedure. Available at: URL: https://cyberleninka.ru/article/n/pravovaya-lingvistika-v-zakonotvorchestvom-protsesse (accessed: 22 -10- 2019)

Kryukova E.A., Kry`zhanovskaya L.A. (2013) Methodological recommendations on the linguistic examination of drafts. Available at: URL: http://www.gosduma.net/analytics/publication-of-legal-department/Metod_lingvo.pdf (accessed: 22-10-2019)

Lundeberg M. (1987) Metacognitive Aspects of Reading Comprehension: Studying Understanding in Legal Case Analysis. *Reading Research Quarterly*, no 4, pp. 407–432. Available at: www.jstor.org/stable/747700 (accessed: 22-10-2019)

Livermore, M., Rockmore D. (eds.) (2019) *Law as Data: Computation, Text, and the Future of Legal Analysis*. Santa Fe: Institute Press, 526 p.

Mikolov T. et al. (2013) Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, vol. 26, pp. 3111–3119.

Owens R., Wedeking J. (2011) Justices and legal clarity: Analyzing the complexity of US Supreme Court opinions. *Law & Society Review*, no 4, pp. 1027–1061.

Polyakov A.V. (2009) Language of legal acts and legal mechanics. A doctrinal and normative commentary to the Federal Law "On State Language of Russia". Saint Petersburg: University, pp. 16–28 (in Russian)

Reynolds R. (2016) Russian Natural Language Processing and Computer-assisted Language Learning: Capturing the benefits of deep morphological analysis in real-life applications. PhD thesis. 172 p. Available at: https://munin.uit.no/bitstream/handle/10037/9685/thesis.pdf. (accessed: 22-10-2019)

Shashek V.V., Kharchenko N. A. (2016) Clarity of legislation in the interpretations of texts of laws. *Molodoy ucheniy*, no 7, pp. 1191–1196 (in Russian)

Smith D., Richardson G. (1999) The Readability of Australia's Taxation Laws and Supplementary Materials: An Empirical Investigation. *Fiscal Studies*, no 3, pp. 321–349.

Stepanov O.A. (2018) Specifying law in the conditions of public practice. *Pravo. Zhurnal Vysshey shkoly ekonomiki*, no 3, pp. 4–23 (in Russian)

Waltl B., Matthes F. (2015) Comparison of Law Texts — An Analysis of German and Austrian Legislation regarding Linguistic and Structural Metrics. Paper presented at the IRIS: Internationales Rechtsinformatik Symposium 2015. Available at: https://wwwmatthes.in.tum.de/pages/1occngdfehma2/Comparison-of-Law-Texts-An-Analysis-of-German-and-Austrian-Legislation-regarding-Linguistic-and-Structural-Metrics (accessed: 22-10-2019)