ISSN 2307-2032 online

ИНСТИТУТ ПРОБЛЕМ ПРАВОПРИМЕНЕНИЯ

ЕВРОПЕЙСКИЙ УНИВЕРСИТЕТ В САНКТ-ПЕТЕРБУРГЕ

АНАЛИТИЧЕСКИЕ ЗАПИСКИ ПО ПРОБЛЕМАМ ПРАВОПРИМЕНЕНИЯ

декабрь 2018

Руслан Кучаков, Денис Савельев

Сложность правовых актов в России

Лексическое и синтаксическое качество текстов

под редакцией Дмитрия Скугаревского

аналитическая записка

ISSN 2307-2032 online
ISSN 2306-6520 print

Руслан Кучаков, Денис Савельев. Сложность правовых актов в России: Лексическое и синтаксическое качество текстов. Под редакцией Дмитрия Скугаревского (Серия «Аналитические записки по проблемам правоприменения»). СПб: ИПП ЕУСПб, 2018. — 20 с.

Институт проблем правоприменения (The Institute for the Rule of Law) создан в 2009 году в составе Европейского Университета в Санкт-Петербурге. Миссия ИПП — содействие судебной реформе и утверждению принципа верховенства права в России. Направления деятельности — проведение научных исследований, публикации и доведение до сведения широкой общественности их результатов, инициация общественных дебатов, выработка стратегических рекомендаций для всех заинтересованных сторон, включая тех, кто принимает решения, а также развитие обучающих программ. Деятельность института поддерживается Сбербанком, Фондом Кудрина по поддержке гражданских инициатив, Российским научным фондом и Европейским университетом в Санкт-Петербурге.

Европейский Университет в Санкт-Петербурге (ЕУСПб) был учрежден в 1994 году и начал свою работу как обучающая аспирантура по социальным наукам в 1996 году. Благодаря высокому профессионализму и уникальному научному потенциалу Европейский университет приобрел репутацию одного из самых динамичных и современных образовательных учреждений страны.

Контакты:

Санкт-Петербург, ул. Гагаринская 3,
Научно-исследовательский центр
«Институт проблем правоприменения»
Тел.: (812) 386 76 12
E-mail: ipp@eu.spb.ru
www.enforce.spb.ru

СОДЕРЖАНИЕ

ОСНОВНЫЕ ВЫВОДЫ | EXECUTIVE SUMMARY

Тексты правовых актов должны быть ясными, доступными по сложности для восприятия и недвусмысленными.

Созданный нами открытый корпус текстов правовых актов позволил проанализировать их сложность методами вычислительной лингвистики. Мы исследовали 458 тысяч текстов правовых актов Российской Федерации и ее субъектов, доступных на Официальном интернет-портале правовой информации.

Сложность текста можно оценивать с разных сторон: по объему и структуре, по сложности терминов, соотнося их со словарями, по синтетическим индексам, по отдельным лингвистическим метрикам. В настоящем обзоре показаны две метрики: лексическое разнообразие (повторяемость одних и тех же слов) и длина зависимых связей в предложении (сложность структуры предложения для чтения). Мы считаем более качественным текст, в котором избегают повторов слов и строят предложения простыми языковыми конструкциями.

В результате проведенного анализа мы пришли к следующим выводам:

- В России наблюдается усложнение текстов правовых актов — падение лексического разнообразия, усложнение структуры предложений. При этом сравнимый корпус текстов СМИ показывает обратные результаты.
- В особенности такое усложнение мы видим в 2014–2017 гг. Сравнимые документы правотворческих органов РСФСР и СССР оказываются проще. Законодательные акты становятся более сложными с течением времени при внесении в них изменений и дополнений.
- Наиболее сложными являются правовые акты Конституционного Суда РФ, а также документы, связанные с финансово-бюджетной сферой.
- Некоторые правовые акты в принципе невозможно оценить с точки зрения сложности или читаемости, поскольку они содержат огромные таблицы чисел или перечни различных сущностей (десятки тысяч страниц).

ВВЕДЕНИЕ

Особенности и качество юридического и, в частности, законодательного языка постоянно становятся предметом обсуждения и анализа в России (Исаков, 2000; Крюкова, Крыжановская, 2013; Шашек, Харченко, 2014 и др.) и мире (например, Owens et al., 2013, DeFriez, 2017 исследовали читаемость судебных решений, Smith, Richardson, 1999 приводят более 25 исследований финансово-налоговых документов). Основным требованием, которое предъявляется к языку текста правового акта, является ясность изложения. Возникло движение за понятный правовой язык[1], однако существует и его критика (Assy, 2011). Вычислительная лингвистика позволила даже провести сравнительное исследование права двух государств (Waltl, Matthes, 2015). Тема сложности, ясности и читаемости законодательного текста, безусловно, является многоаспектной и неисчерпаемой. Некоторые свойства текста в целом, слов и предложений, из которых он состоит, можно измерить количественно. Опубликование правовых актов в электронной форме открывает возможность изучения их текстов методами вычислительной лингвистики. На основе опубликованных на Официальном интернет-портале правовой информации правовых актов составлен корпус текстов в специализированном формате, пригодный для количественного анализа (Савельев, 2018). Этот корпус покрывает все доступные на портале документы по 2017 год (458 тысяч текстов)[2]. Таким образом, настоящее исследование[3] охватывает значительный объем правовых актов.

В вычислительной лингвистике разработаны метрики читаемости текста, которые универсально применимы к любым текстам, позволяют количественно выявить свойства используемого в тексте языка и оценить таким образом его сложность (Pitler, Nenkova, 2008; Feng et al., 2010; для русского языка см. Reynolds, 2014).

Из различных метрик составляются формулы, которые в целом могут ранжировать тексты по сложности восприятия с учетом возраста и уровня образования, такие как индекс удобочитаемости Флеша (Flesch, 1948) и др., которые на время стали стандартом индустрии. Однако в предсказании сложности неакадемического текста для взрослого читателя формулы, представленные в прошлом, уступают новым методикам обработки текстов на естественном языке (Crossley et al., 2017).

В связи с этим каждый документ был оценен по 24 метрикам, которые в предшествующих исследованиях русскоязычных текстов (Reynolds, 2014) дали наилучший результат. Полученные данные были обобщены с помощью метода главных компонент[4]. По итогам мы отобрали две взаимодополняющие метрики, описанные ниже. Они характеризуют повторяемость слов и сложность построения структуры предложений. В настоящем исследовании не рассматривается смысловая сложность используемых понятий. Анализируются метрики синтаксической и лексической сложности текста, которые учитывают только количественные значения, выражающие формальное описание структуры взаимосвязей частей речи в предложениях.

На данных используемого корпуса для каждого правового акта мы рассчитываем две избранные метрики сложности текста: лексическое разнообразие и расстояние между зависимыми словами в предложении.

---

[1] В России это движение получило развитие в проекте «Простым языком» НП «Информационная культура». URL: http://ru.readability.io.

[2] Техническое описание процесса подготовки и анализа данных представлено в Приложении.

[3] Публикация подготовлена в рамках научного проекта No 17-18-01618, поддержанного Российским научным фондом.

[4] Метод главных компонент, англ. Principal Component Analysis (PCA), является одним из базовых методов понижения размерности данных, он позволяет «сжать» информацию, содержащуюся в большом наборе объясняющих характеристик, до нескольких новых переменных (компонент). Метод главных компонент также позволяет оценить величину вклада каждой исходной переменной в создание новой.

Лексическое разнообразие

Одна из наиболее простых метрик читаемости текста — коэффициент лексического разнообразия, который позволяет сравнить тексты по уровню вариативности языка. Определим коэффициент TTR (англ. Type-Token Ratio) как:

$$TTR = \frac{\text{число уникальных токенов}}{\text{общее число токенов}} .$$

Иными словами, это отношение числа уникальных токенов (слов, чисел и т. д.)[5] к их общему количеству. TTR измеряется от 0 до 1. Он равен единице, если все слова в тексте различны, и стремится к нулю, если в тексте много повторов одних и тех же слов. Эта метрика достаточно давно используется в вычислительной лингвистике и ее применение имеет положительные и отрицательные стороны (Richards, 1987), что требует определенной интерпретации с нашей стороны.

В обычных условиях, например когда речь идет об обучении языку, текст можно считать более сложным для восприятия, если он содержит множество разнообразных слов. Проще будет тот текст, в котором такое разнообразие меньше. Однако возвращаясь к текстам правовых актов, мы наблюдаем иную картину: 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. В тексте, тем более в юридическом, нет необходимости добиваться абсолютного разнообразия слов (TTR=1), однако следует избегать излишних повторов.

Расстояние между зависимыми словами в предложении (maxDepLen)

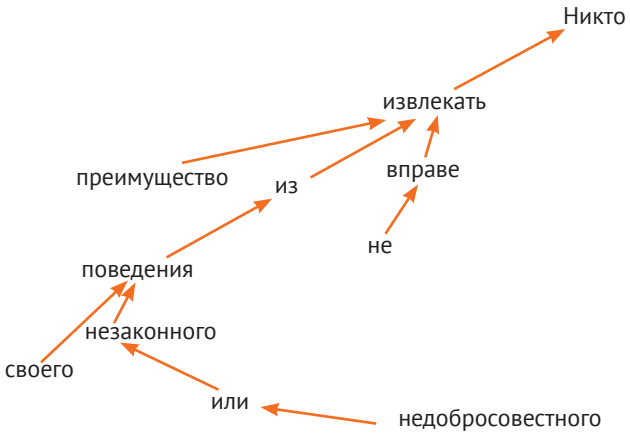
Длина расстояний между зависимыми частями речи в предложении (англ. dependency length, DepLen) может показать сложность структуры предложения. Эта метрика говорит о том, через сколько слов человек должен «перескочить» в предложении, чтобы считать смысл зависимых друг от друга слов. Таким образом, она показывает общее усложнение текста независимо от смысла употребленных слов. Максимальное значение среди всех расстояний в предложении можно обозначить как maxDepLen (Reynolds, 2014).

Благодаря развитию технологий обработки текстов на естественном языке возможно распознавать структуру предложений и элементов в них. Для каждого слова можно определить, что это за часть речи, в какой форме стоит, какова начальная форма слова, с какими другими словами оно связано (например, к какому существительному относится прилагательное). Представить структуру зависимостей слов в каждом предложении текста можно в виде дерева подчинений, где его ветви — зависимости подчиненных друг другу частей речи. Зависимость устанавливается исходя из правил, на основе которых слова соединяются в правильно построенные предложения. Например, существительное связывается с прилагательным определительным отношением (подробнее см. Боярский, 2013). На рисунке 1 в качестве примера приведена структура зависимостей (дерево подчинения) одного предложения из текста правового акта.

---

[5] Под токеном (текстоформой) понимается выделенный в текстовом потоке минимальный фрагмент для последующего анализа (слово, число, знак препинания и т. д.). Уникальными будут считаться токены, встретившиеся лишь однажды в тексте. Выделение токенов может отличаться в зависимости от используемого программного обеспечения. В настоящем исследовании используется TTR без лемматизации, т. е. приведения к начальной форме слова. Например, «закон» и «закона» будут считаться разными токенами. Тем не менее, и TTR с использованием лемматизации токенов показывает аналогичные результаты.

Рис. 1 Дерево зависимостей предложения из п. 4 ст. 1 ГК РФ

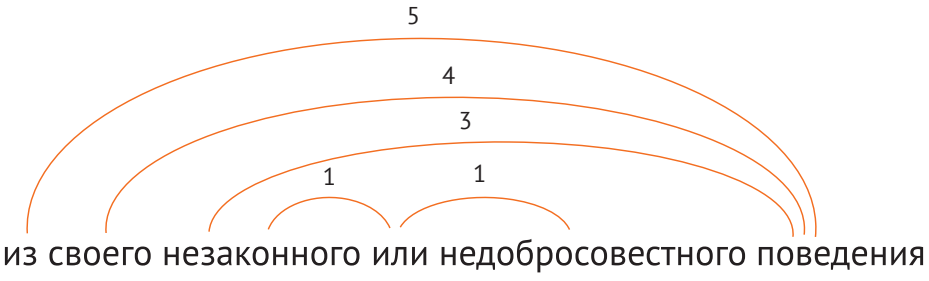


На рисунке представлена в виде дерева зависимых частей речи ч. 4 ст. 1 ГК РФ: «Никто не вправе извлекать преимущество из своего незаконного или недобросовестного поведения».

Однако не все предложения правовых актов и решений высших судов, публикуемых на официальном портале, так же просты, как приведенное на рисунке 1, — нередки предложения длиной более 100 слов, в которых много длинных «ветвей».

Показанные выше зависимости могут быть между соседними словами в предложении, а могут быть между словами в разных концах предложения. Оценить сложность предложения можно, определив максимальное число слов, лежащих в предложении между двумя зависимыми словами. На рисунке 2 представлена иллюстрация подсчета этой метрики.

Рис. 2 Схема расчета величины максимальной длины расстояний между зависимыми частями речи в предложении (maxDepLen)



DepLen = 14, maxDepLen = 5

В случае когда зависимости в предложении пересекаются или «накрывают» одна другую (высокий maxDepLen), текст становится труднее для восприятия, так как для считывания смысла требуется преодолеть несколько обособленных речевых конструкций. Напротив, если зависимые члены предложения идут в предложении последовательно (низкий maxDepLen), как звенья цепи, то текст становится проще. MaxDepLen исчисляется от единицы до бесконечности.

Вычисление значений избранных метрик для документов

Существует возможность вычислять различные метрики, в том числе описанные выше, на уровне предложений и на уровне документа в целом, а затем использовать средние или максимальные значения. В настоящем исследовании мы использовали значения TTR как среднее для документа без разделения текста на предложения или отдельные фрагменты — считая число и уникальность токенов в рамках всего текста. Что касается максимальной длины расстояний между зависимыми частями речи в предложении (maxDepLen), для каждого конкретного текста взято одно значение, которое является максимальным для всех предложений текста. Таким образом, мы оцениваем как текст в целом, так и одно наиболее сложное предложение этого текста. Далее вычисляются средние значения указанных метрик по всем текстам за год.

Усложнение текстов законодательных актов на примерах конкретных документов

При чтении правовые акты первых лет современной российской государственности кажутся гораздо более простыми для восприятия, чем недавно принятые документы. Для примера можно взять Гражданский кодекс РФ. Часть первая была принята в 1994 г. и содержит общие положения гражданского законодательства. В 2006 г. была принята часть четвертая, которая содержит нормы об интеллектуальных правах. В ее тексте мы увидим существенно более массивные статьи, длинные предложения, сложные конструкции. Возникает вопрос, оправданно ли для регулирования одного из институтов (интеллектуальных прав) были использованы более сложные предложения, чем для регулирования важнейших начал законодательства, например положений о собственности, о лицах, о договорах, или же это связано с неустоявшимся понятийным аппаратом и несложившимся комплексом норм об интеллектуальных правах? Но даже если считать, что общая часть кодекса может быть проще особенной части, сравнение первоначальной и действующей (с учетом изменений и дополнений) редакций первой части показывает существенное увеличение и усложнение текста. Подчеркнем еще раз, что речь идет именно о лексической и синтаксической, а не о семантической сложности текста.

Табл. 1 Метрики читаемости редакций ГК РФ и закона об образовании

Документ	*Число токенов*	*TTR*	*maxDepLen*
ГК РФ, ч. 1, 1994	45 101	0,12	71
ГК РФ, ч. 4, 2006	52 081	0,09	78
ГК РФ, ч. 1, действ. ред.	80 746	0,09	73
Закон «Об образовании», 1992	11 346	0,22	53
ФЗ «Об образовании», действ. ред.	54 777	0,09	201

Еще один пример — Закон РФ «Об образовании» в первоначальной редакции (1992 г.) и вновь принятый в 2012 г. Федеральный закон «Об образовании в Российской Федерации» (действующая редакция). Мы можем наблюдать четырехкратное увеличение максимального расстояния зависимых слов в предложении, увеличение объема документа примерно в 5 раз и снижение лексического разнообразия. В действующей редакции закона десятки предложений достигают почти 100 токенов в длину, а максимально длинное предложение состоит из 390 токенов.

Изменение значений избранных метрик законодательства и СМИ во времени

Показанное выше изменение метрик читаемости одного документа во времени дает основания для более масштабного сравнения документов, принятых в разное время. Для того чтобы сделать вывод об изменениях метрик во времени, мы рассчитали среднегодовые значения по вновь принятым в каждый год правовым актам.

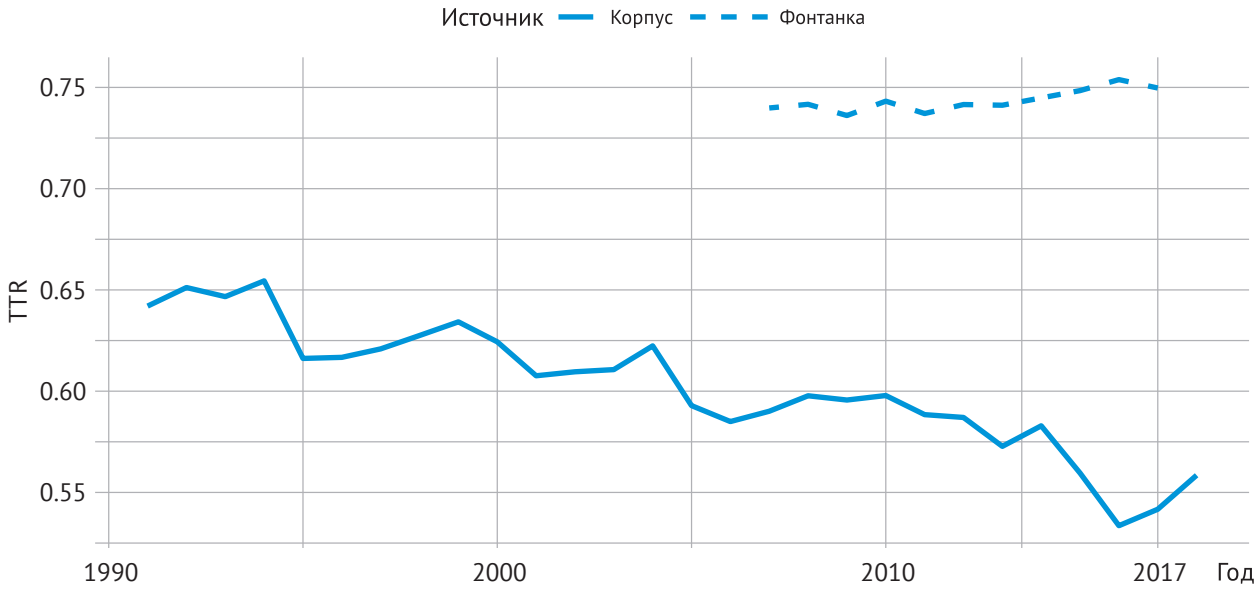
Может возникнуть предположение, что возможное усложнение законодательных текстов является объективным явлением и следует за усложнением языка, описывающего общественно-политическую жизнь. Тогда вопрос о том, соответствуют ли изменения сложности законодательства общему изменению общественно-политических текстов, можно исследовать путем сравнения среднегодовых значений аналогичных метрик доступного по времени корпуса публикаций СМИ. Для этой цели мы использовали публикации интернет-издания «Фонтанка.ру»[6] за 2007–2017 гг. из открытого корпуса текстов «Тайга» (Shavrina, Shapovalova, 2017). Такие сравнения отражены на рисунках 3 и 4. Сравнение линий на рисунке 3 показывает — тренды разнонаправлены; масштаб изменения у текстов «Фонтанки» существенно меньше, а разнообразие слов — существенно выше, чем в законодательстве; после 2014 г. темп изменений усиливается в обоих корпусах. Снижение среднегодового значения TTR в корпусе правовых актов свидетельствует о том, что чаще повторяются одни и те же слова. По нашим наблюдениям, основанным на данных частотного употребления токенов, такое увеличение повторов достигается за счет многократного упоминания выражения «Российская Федерация» и различных субъектов права. При этом на корпусе СМИ аналогичного увеличения повторов мы не наблюдаем, наоборот, есть слабый рост лексического разнообразия. Так же, как и в случае с TTR, мы видим разнонаправленные тренды по максимальной длине дерева зависимостей слов в предложении (переменная maxDepLen, рис. 4). MaxDepLen ухудшается (растет) в текстах правовых актов и улучшается в текстах СМИ. Из рисунка 4 следует, что в правовых текстах с течением времени образуются все более длинные расстояния между зависимыми словами.

В отличие от тенденции, которую мы увидели в российском законодательстве, в 37 языках разных стран отмечено снижение средней длины зависимых частей речи (Futrell et al., 2008).

---

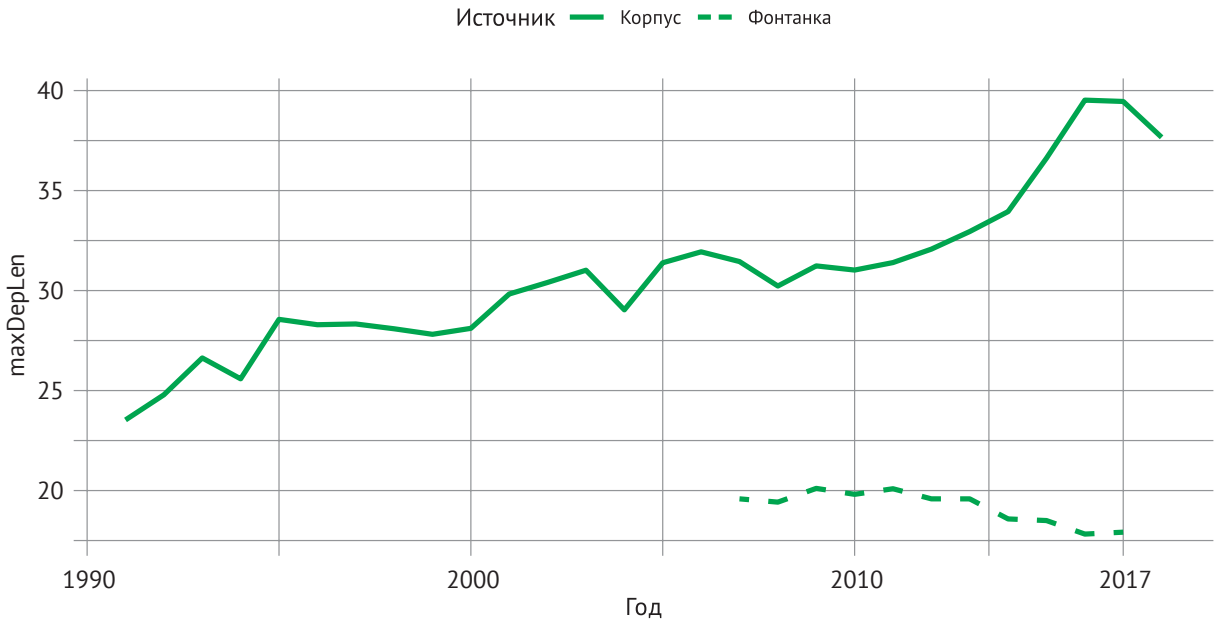
[6] Крупное региональное интернет-издание, существующее с 1999 г. URL: https://www.fontanka.ru/.

Рис. 3 Среднегодовые значения TTR в законодательстве и СМИ



На оси X отложен год принятия правового акта. На оси Y — среднегодовые значения TTR. TTR принимает значения от 0 до 1, где 1 — максимальное разнообразие; чем ниже значение TTR, тем сложнее текст для восприятия. Сплошная линия — среднегодовые показатели TTR, рассчитанные на данных выборки всех правовых актов органов власти федерального уровня из корпуса (Савельев, 2018). Использованы данные только документов таких видов, которые встречались 10 и более раз. Пунктирная линия — среднегодовые показатели TTR, рассчитанные на данных по публикациям «Фонтанка.ру» от 2007 г. (Shavrina, Shapovalova, 2017). Тексты публикаций, содержащие большие таблицы и перечни, были исключены из рассмотрения.

Рис. 4 Среднегодовые значения maxDepLen в законодательстве и СМИ



На оси X отложен год принятия правового акта. На оси Y — среднегодовые значения maxDepLen. MaxDepLen — это максимальное число токенов в тексте между подлежащим и сказуемым. MaxDepLen принимает только целочисленные значения и изменяется от нуля до бесконечности; чем больше значение maxDepLen, тем сложнее текст. Сплошная линия — среднегодовые показатели maxDepLen, рассчитанные на данных выборки всех правовых актов органов власти федерального уровня из корпуса (Савельев, 2018). Использованы данные только документов таких видов, которые встречались 10 и более раз. Пунктирная линия — среднегодовые показатели maxDepLen, рассчитанные на данных по публикациям «Фонтанка.ру» от 2007 г. (Shavrina, Shapovalova, 2017). Тексты публикаций, содержащие большие таблицы и перечни, были исключены из рассмотрения.

Анализ тенденции к усложнению правовых актов

Среднегодовые значения метрик говорят о явной тенденции к усложнению правовых текстов в России. Однако могут возникнуть возражения, что средний показатель дает некорректную картину. При подсчете среднегодовых значений были усреднены метрики сложности документов разных органов власти, разной тематики и разной длины. Число принятых правовых актов в разные года также различается. Все эти возражения могут представить среднюю слабо пригодным аналитическим инструментом для оценки тенденции.

Регрессионный анализ позволяет решить эту проблему. С его помощью можно оценить сложность текстов в каждый год, при неизменных во времени[7] значениях органа власти, типа документа и длины нормативного акта. Объединив предсказанные значения метрик линиями, мы получили тенденцию (см. рис. 5), свободную от влияния указанных переменных.

Показанная на рисунке 5 динамика соотносима с приведенными выше графиками среднегодовых значений отдельных метрик, что показывает обоснованность полученных в них сравнений и их независимость от иных факторов, кроме измеряемых. Результаты регрессионного анализа позволяют утверждать, что текст правового акта, выпущенный средним органом власти по средней теме со средним названием в 2000 г., будет значимо проще для понимания, чем текст, выпущенный в 2018 г.

Сравнение избранных метрик по документам различных органов власти

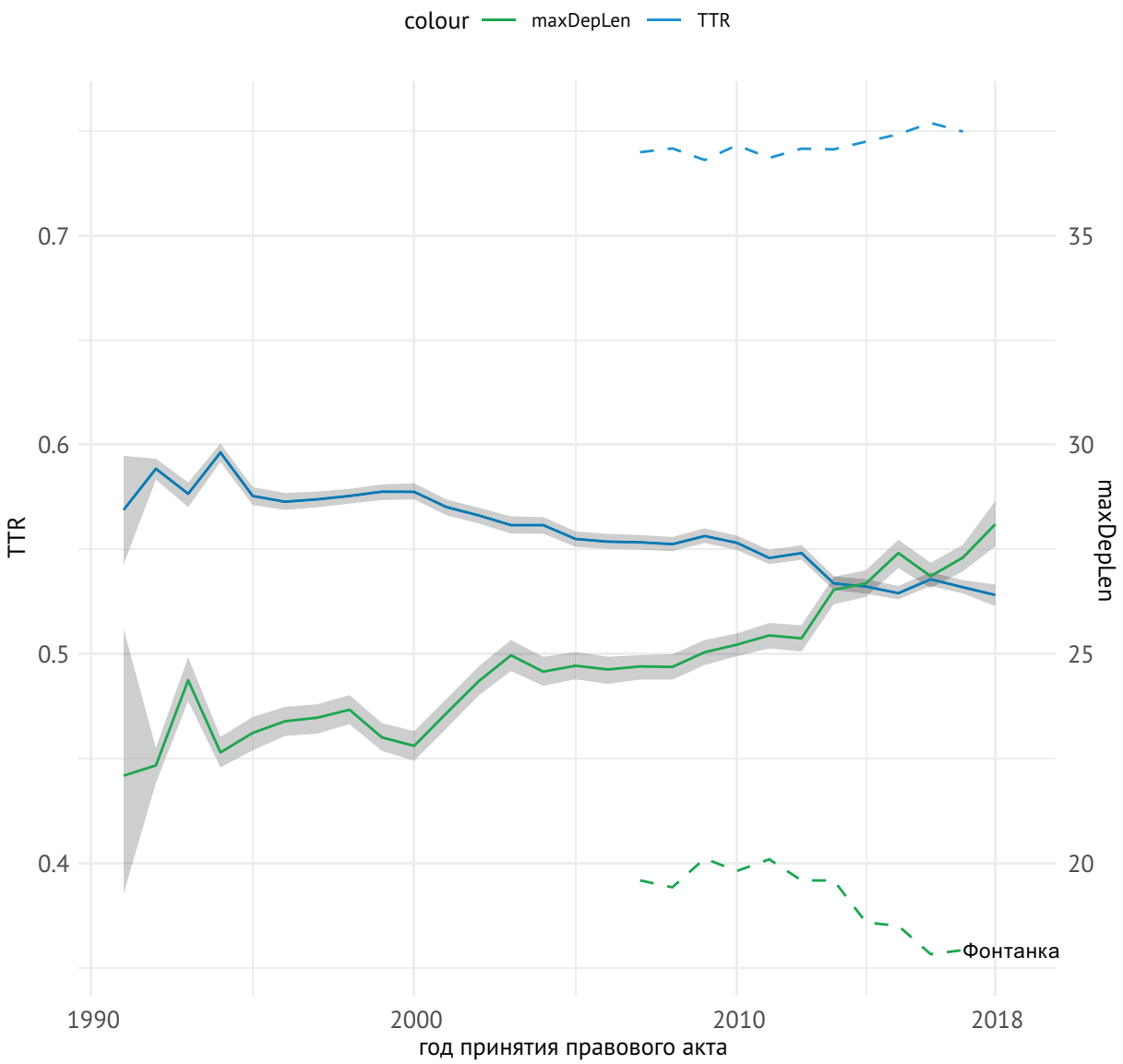
Сравнивая одновременно два приведенных выше показателя — TTR и maxDepLen — для разных видов документов, принятых различными правотворческими органами, мы можем увидеть, документы каких органов власти содержат одновременно много повторов и сложные предложения, или, наоборот, содержат мало повторов и простые предложения, или иную промежуточную вариацию. Такая картина представлена на рисунке 6.

Представленная на рисунке 6 диаграмма показывает, что документы Конституционного Суда РФ, а также документы ведомств, связанных с финансами и бюджетом, отличаются от большинства остальных документов одновременно и малым лексическим разнообразием (TTR), и длинными связями внутри предложения (maxDepLen). Напомним, что табличные по сути документы и перечни из анализа исключены. В настоящем исследовании мы не касаемся причин отличий показателей читаемости документов Конституционного Суда РФ, которые исследовались ранее (Дмитриева, 2017). В противоположной стороне от основных документов находятся различные документы органов власти СССР и РСФСР, доступные в корпусе (Савельев, 2018). Это показывает, что такие документы отличались большим разнообразием, одновременно с меньшей структурой зависимостей слов в предложениях. Последнее также соотносится с приведенными выше хронологическими графиками изменения метрик.

---

[7] Данная процедура называется вычислением предельных эффектов. В работе используется подход marginal effects at the means: изменение зависимой переменной, например метрик читаемости, оценивается при изменении выбранной переменной, в нашем случае года принятия правового акта, и фиксации всех остальных регрессоров на средних значениях.

Рис. 5 Предсказанные годовые значения TTR и maxDepLen при фиксации эффекта года, учреждения и тематики



На оси X отложен год принятия правового акта. На осях Y отложено две шкалы, соответствующие двум разным метрикам читаемости текста. Шкала слева — TTR — принимает значения от 0 до 1, где 1 — максимальное разнообразие; чем ниже значение TTR, тем сложнее текст для восприятия. Шкала справа — maxDepLen — это максимальное число токенов в тексте между словами, находящимися в синтаксической зависимости. MaxDepLen принимает только целочисленные значения и изменяется от нуля до бесконечности; чем больше значение maxDepLen, тем сложнее текст. Сплошные линии отражают предсказанные значения соответствующих метрик по годам из регрессий данной метрики на год принятия правового акта, принявший документ орган власти и тематику документа. Контроль тематики осуществляется с помощью включения регрессоров-токенов из названий документов. Использованы данные только документов таких видов, которые встречались 10 и более раз. Серые области отражают доверительные интервалы, построенные с помощью бутстрепа с 1000 репликаций и покрывают 95% значений переменной. Пунктирные линии — среднегодовые показатели TTR и maxDepLen, рассчитанные на данных по публикациям «Фонтанка.ру» от 2007 г. (Shavrina, Shapovalova, 2017). Пунктирные линии представлены для задания масштаба изменений. Тексты публикаций, содержащие большие таблицы и перечни, были исключены из рассмотрения.

Рис. 6 Предсказанные значения TTR и maxDepLen для различных видов документов при фиксации эффекта года, принявшего документ органа власти и тематики

На осях X и Y отложены предсказанные значения TTR и maxDepLen, соответственно, из регрессии на год принятия правового акта, принявший документ орган власти и тематику документа. Контроль тематики осуществлялся с помощью включения регрессоров-токенов из названий документов. Размер точек соответствует количеству документов в корпусе. Красной обводкой выделены органы власти СССР и РСФСР. На рисунке отражены только те органы власти, чей последний нормативный акт был принят до 1993 г. или после 2016 г. включительно, а общее количество принятых актов превышает десять. Это ограничение сохраняет в выборке советские органы власти, но исключает упраздненные к 2016 г. правотворческие органы. Всего на рисунке изображено 226 уникальных пар типа документа и государственного органа, например Распоряжение Президента РФ и Указ Президента РФ представлены отдельными точками. Для облегчения восприятия рисунка подписаны только отдельные сочетания.

Эксцессы увеличения объема правовых актов

Невозможно не обратить внимание на то, что в некоторых документах длина дерева зависимостей (и соответственно, длина предложения) превышает тысячи слов. Такие документы явно непригодны для оценки с точки зрения читаемости предложений. Исследуя их, мы нашли правовые акты огромных объемов. Причем объем документов не всегда оправдан содержанием. Самым первым документом объемом более 1000 страниц, официально опубликованным в электронной форме на официальном интернет-портале правовой информации pravo.gov.ru, был бюджет на 2012 г. (4000 страниц). Бюджеты и отчеты об их исполнении, федеральные целевые программы всегда превышают по количеству страниц многие правовые акты, но это далеко не все большие документы. В том же 2012 г. был опубликован протокол о присоединении России к Всемирной торговой организации (5500 страниц, часть — таблицы, часть — на английском языке).

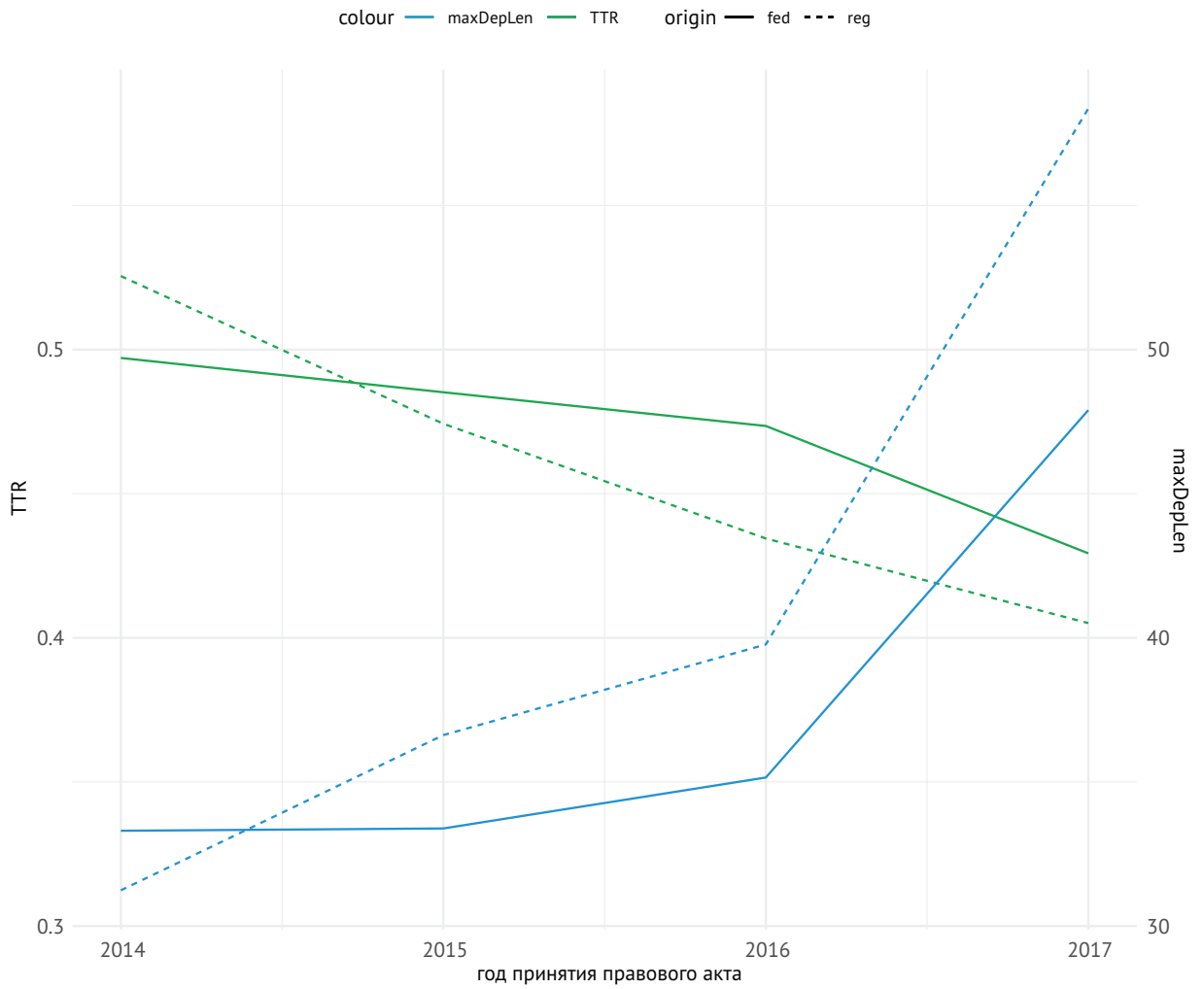
Существуют правовые акты, которые в принципе невозможно оценить с точки зрения сложности или читаемости, поскольку они содержат огромные таблицы чисел или перечни различных сущностей. Например, в 2013 г. постановлением правительства РФ № 534 была утверждена таблица из 1600 страниц, в которой содержатся только координаты точек на карте, обозначающие границы Сочинского национального парка. Региональное законодательство содержит еще больших подобных документов, чем федеральное. Так, закон Новосибирской области на 9800 страницах излагает координаты границ муниципальных образований. Самые большие документы, 15 000–35 000 страниц, — это данные оценки кадастровой стоимости объектов и региональные программы капитального ремонта, где в таблицах перечислен каждый дом в регионе. Например, соответствующее постановление правительства Тульской области с приложением состоит из 29 500 страниц. При этом приложение формально является неотъемлемой частью этого документа. На его распечатку (а это необходимо для его подписания) потребуется 59 пачек бумаги общим весом примерно 169 кг. С учетом того, что закон опубликован в графической форме, где поиск по тексту невозможен, практической пользы от такого официального опубликования немного.

Такие документы поднимают системную проблему: формально руководитель органа власти, поставивший свою подпись под таким документом, должен был прочитать, осознать и своей волей утвердить документ. При стандартной скорости чтения это потребовало бы от подписавшего документ лица не менее трех рабочих месяцев, когда он не должен заниматься ничем, кроме такого чтения. Но на самом деле он утверждает уже не текст правового акта, а распечатанный результат работы по наполнению соответствующей базы данных. Необходимо найти механизмы, позволяющие сократить такие огромные документы за счет вынесения из них данных в электронные ресурсы.

Сложность регионального законодательства

Официальный интернет-портал правовой информации www.pravo.gov.ru начал официальное опубликование правовых актов федеральных органов власти в виде графических образов страниц в 2011 году. Официальное опубликование правовых актов субъектов федерации в электронной форме на портале осуществляется с 2014 г. Разные субъекты федерации подключались к этой работе не одновременно, к концу 2017 года на портале представлены правовые акты 47 субъектов федерации. К моменту создания исследуемого корпуса текстов в нем представлено около 200 тыс. таких правовых актов, прошедших распознавание. Это позволило нам оценить представленные выше метрики сложности текстов в региональном разрезе на имеющемся объеме текстов и сравнить с федеральным законодательством. На рисунке 7 видно, что законодательство субъектов федерации также усложняется.

Рис. 7 Изменение метрик TTR и maxDepLen правовых актов, принятых органами власти Российской Федерации и её субъектов



На оси X отложен год принятия правового акта. На оси Y отложено два показателя читаемости: слева — TTR, справа — maxDepLen. Сплошные линии отражают значения метрик правовых актов федеральных органов власти, пунктирные — субъектов федерации. Линии объединяют предсказанные значения метрик по годам из регрессии читаемости на год принятия правового акта, орган, принявший документ, и тематику документа. Контроль тематики осуществлялся с помощью включения регрессоров-токенов из названий документов. Регрессия правовых актов субъектов федерации контролировалась на регион, контроль на орган власти исключен из-за возможной мультиколлинеарности. Использованы данные только документов таких видов, которые встречались 10 и более раз. Данные по документам субъектов федерации вычислены на текстах, прошедших оптическое распознавание графических страниц (Савельев, 2018), федеральных — из текстов, полученных без распознавания. Региональные правовые акты доступны за 2014–2017 гг., 135 тыс. документов.

Представление законодательства субъектов федерации на диаграмме рассеяния, которая показывает разницу в лексическом разнообразии и сложности предложений документов разных субъектов, дает представление об общей картине качества регионального законодательства.

Рис. 8 Предсказанные значения TTR и maxDepLen для документов субъектов федерации при фиксации эффекта года и тематики

На осях X и Y отложены предсказанные значения TTR и maxDepLen, соответственно, из регрессии указанной метрики на год принятия правового акта, регион и тематику документа. Контроль тематики осуществлялся с помощью включения регрессоров-токенов из названий документов. Из данных исключены документы таких видов и органов власти, пересечения которых встречаются в корпусе менее десяти раз. Данные по документам субъектов федерации вычислены на текстах, прошедших оптическое распознавание графических страниц (Савельев, 2018), федеральных — из текстов, полученных без распознавания. Региональные правовые акты доступны за 2014–2017 гг., 135 тыс. документов. Для сохранения читаемости рисунка названия регионов сокращены.

Из представленной на рисунке 8 картины можно сделать следующие выводы. Некоторые регионы часто допускают случаи создания документов с большой максимальной длиной зависимостей, что выражается в достаточно высокой средней (например, Архангельская, Иркутская область, Краснодарский край), тогда как иные регионы (например, Дагестан, Удмуртия, Осетия, Приморский край) создают документы из более простых предложений. При этом есть регионы, которые допускают много повторов одинаковых слов (Архангельская, Воронежская области). По Чеченской республике невозможно сделать обоснованного вывода ввиду малого количества документов.

СПИСОК ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

Боярский К. К. Введение в компьютерную лингвистику : учебное пособие. СПб. : НИУ ИТМО, 2013. 72 с.

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

Исаков В. Б. Язык права // Юрислингвистика-2 : Русский язык в естественном и юридическом бытии : межвуз. сб. науч. тр. / под. ред. Н. Д. Голева. Барнаул, 2000. С. 72–89.

Крюкова Е. А., *Крыжановская Л. А.* Методические рекомендации по лингвистической экспертизе законопроектов [Электронный ресурс]. М., 2013. URL: http://www.gosduma.net/analytics/publication-of-legal-department/Metod_lingvo.pdf (дата обращения: 31.10.2018).

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.

Ткаченко Н. В. Статистический анализ федерального законодательства [Электронный ресурс]. М., 2017. URL: https://csr.ru/wp-content/uploads/2017/02/Issledovanie_TSSR_statistika-po-zakonoproektam.pdf (дата обращения: 31.10.2018).

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. № 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. Т. 38, № 3. Pp. 376–404.

Brian R. Type-Token Ratios: What do they really tell us? // Journal of Child Language. 1987. Т. 14, № 2. Pp. 201–209.

Crossley S. A. et al. Predicting text comprehension, processing, and familiarity in adult readers : New approaches to readability formulas // Discourse Processes. 2017. Т. 54. № 5–6. Pp. 340–359.

DeFriez B. M. Toward a Clearer Democracy: The Readability of Idaho Supreme Court Opinions as a Measure of the Court's Democratic Legitimacy : дис. – University of Idaho, 2017.

Feng L., *Jansche M.*, *Huenerfauth M.*, Elhadad N. A comparison of features for automatic readability assessment. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA. 2010. Pp. 276–284.

Flesch, R. A new readability yardstick // Journal of Applied Psychology. 1948. № 32. Pp. 221–233.

Futrell R., *Mahowald K.*, *Gibson E.* Large-scale evidence of dependency length minimization in 37 languages. Proceedings of the National Academy of Sciences of the United States of America PNAS. 2015. № 112 (33). Pp. 10336–10341.

Gildea D., *Temperley D.* Do Grammars Minimize Dependency Length? // Cognitive Science. 2010. № 34. Pp. 286–310.

Karimi S. Word Order and Scrambling / Blackwell Publishing Ltd, 2008. 385 p.

Kettunen K. Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? // Journal of Quantitative Linguistics. 2014. T. 21, № 3. Pp. 223–245.

Owens R.J., *Wedeking J.*, *Wohlfarth P.C.* How the Supreme Court alters opinion language to evade congressional review // Journal of Law and Courts. 2013. T. 1, № 1. Pp. 35–59.

Pitler E., *Nenkova A.* Revisiting Readability: A Unified Framework for Predicting Text Quality. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu. 2008. Pp. 186–195.

Reynolds R. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, San Diego, CA: 16 June 2016. In: Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications. Pp. 289–300.

Reynolds R. Russian Natural Language Processing and Computer-assisted Language Learning : Capturing the benefits of deep morphological analysis in real-life applications : PhD thesis / Tromsø : Universitet i Tromsø. 2016. URL: https://munin.uit.no/bitstream/handle/10037/9685/thesis.pdf.

Shavrina T., Shapovalova O. To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser. Proceedings of the international conference «Corpus Linguistics — 2017», Saint Petersburg, 2017.

Smith D., *Richardson G.* The Readability of Australia's Taxation Laws and Supplementary Materials: An Empirical Investigation // Fiscal Studies. 1999. T. 20, № 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 the IRIS: Internationales Rechtsinformatik Symposium, Salzburg, Austria, 2015 [Электронный ресурс]. URL: 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: 31.10.2018).

ПРИЛОЖЕНИЕ 1.
ОПИСАНИЕ РАБОТЫ С ДАННЫМИ

Общая характеристика корпуса

Предметом исследования стал корпус текстов вновь принятых правовых актов[8], который сформирован на основе документов, доступных на Официальном интернет-портале правовой информации[9]. Часть документов на портале прошла официальное электронное опубликование и их тексты получены нами после оптического распознавания сканов графических страниц (OCR). Остальная часть получена из HTML-текстов раздела «Законодательство России» портала parvo.gov.ru.

В корпус текстов включены только первоначальные редакции правовых актов на момент принятия, а также последующие акты об изменениях. Консолидированные версии документов с внесенными изменениями не использовались (кроме примера из раздела 4) во избежание двойного подсчета.

Использованный в настоящем исследовании объем корпуса текстов — 458 884 документов за 1991–2017 гг. Из них 203 828 документов не требовали OCR. В этой части объем обработанного корпуса в токенах — 126 млн. Всего в корпусе более 600 млн токенов. Статистика по видам текстов представлена в таблице 1.

Табл. 1 Количество токенов в текстах

	Всего токенов[10]	*Уникальных токенов*	*Уникальных лемм*[11]
Тексты без OCR	126 134 021	879 402	527 246
Тексты после OCR	497 440 264	7 106 928	5 860 010

Обработка и подготовка текста

Для увеличения точности подсчета метрик тексты, которые не требовали OCR, прошли предварительную обработку. Во-первых, из текстов были удалены таблицы (ввиду того что в таблицах сложно выделить предложения целиком) и части, которые представляют собой формы для заполнения (большую часть текста в них занимают знаки подчеркивания).

---

8 Russian Law as Open Data. URL: https://github.com/irlcode/RusLawOD.

9 Официальный интернет-портал правовой информации. URL: http://www.pravo.gov.ru.

10 Пояснения относительно понятия токенов указаны во введении.

11 Лемма — слово, приведенное в начальную форму. Например, для существительных это именительный падеж единственного числа. Лемматизация осуществляется в ходе морфосинтаксической разметки.

Во-вторых, в документах можно выделить собственно текст и служебные части: начало с метаданными, окончание с подписью и внутри текста — заголовки и подзаголовки структурных элементов. Все эти части из документов были удалены так, чтобы оставался только текст структурных элементов.

Очень важной для подсчета длины предложения является сегментация по предложениям. В документах часто встречаются списки различных сущностей. В этом случае предложениями считаются начало списка до двоеточия и каждый элемент списка отдельно. Также для качественной сегментации в предложениях выделялись сокращения с точкой, даты. Из документов также удалены ссылки на источники опубликования правовых актов.

Морфосинтаксическая разметка

Морфосинтаксическая разметка корпуса — это определение свойств каждого токена. Она производится путем создания файлов в специальном формате[12], в которых текст разбит сначала на предложения, затем на токены, затем для каждого токена определен ряд лингвистических свойств (начальная форма слова, использованная форма слова, место по отношению к главному и т. п.). Данная работа проведена автоматизированным образом с помощью пакета программ ru-syntax[13], подготовленного кафедрой компьютерной лингвистики НИУ ВШЭ при использовании машинного обучения на подготовленных текстах[14]. В данном пакете связаны между собой морфологический анализатор MyStem, программа частеречной разметки TreeTagger и анализатор зависимостей MaltParser.

---

[12] См., например: URL: http://universaldependencies.org/format.html.

[13] См. URL: https://github.com/tiefling-cat/ru-syntax.

[14] См. URL: https://www.hse.ru/ma/ling/.

ИНСТИТУТ ПРОБЛЕМ ПРАВОПРИМЕНЕНИЯ

Сотрудники

Вадим Волков *научный руководитель*	доктор социологических наук, PhD (Cambridge University), профессор социологии права им. С.А. Муромцева Европейского университета в Санкт-Петербурге, автор книги «Силовое предпринимательство: экономико-социологический анализ» (2005)
Кирилл Титаев	кандидат социологических наук, специалист по эмпирическим исследованиям права и правоприменения
Арина Дмитриева	социолог, экономист, специалист по экономическому анализу права
Мария Шклярук	юрист, LL.M. (Hamburg), кандидат экономических наук, специалист по проблемам правоохранительной деятельности, сравнительному правоведению
Дмитрий Скугаревский	PhD, экономист, специалист по судебной статистике
Ирина Четверикова	юрист, социолог, специалист по судебной статистике и криминологии
Екатерина Моисеева	кандидат социологических наук, специалист по социологическому анализу рынков и социологии профессий
Тимур Бочаров	юрист, социолог, специалист по гражданскому и арбитражному процессу
Алексей Кнорре	социолог, программист
Владимир Кудрявцев	политолог, специалист по уголовной политике
Денис Савельев	кандидат юридических наук, специалист по информационному праву и интеллектуальным правам
Дарья Кузнецова	юрист, специалист в антимонопольном праве и концессионных соглашениях
Руслан Кучаков	экономист, политолог, специалист по контрольно-надзорной деятельности

Административный директор — **Мария Батыгина** Администратор — **Маргарита Михно**

Наши книги

- *Э. Панеях, К. Титаев, М. Шклярук*. Траектория уголовного дела: институциональный анализ. СПб., 2018. — 476 с.
- *Т. Бочаров, Е. Моисеева*. Быть адвокатом в России: социологическое исследование профессии. СПб., 2016. — 278 с.
- *К. Титаев*, *М. Шклярук*. Российский следователь: призвание, профессия, повседневность. М., 2016. — 192 с.
- *В. Волков*, *А. Дмитриева*, *М. Поздняков*, *К. Титаев*. Российские судьи: социологическое исследование профессии. М., 2016. — 272 с.
- Обвинение и оправдание в постсоветской уголовной юстиции. М., 2015. — 320 с.
- Право и правоприменение в зеркале социальных наук: хрестоматия современных текстов. М., 2014. — 568 с.
- По ту сторону права: законодатели, суды и полиция в России. М., 2014. — 331 с.
- Как судьи принимают решения: эмпирические исследования права. М., 2012. — 368 с.
- Право и правоприменение в России: междисциплинарные исследования. М., 2011. — 317 с.

Наши серии

- Аналитические записки по проблемам правоприменения
- Аналитические обзоры по проблемам правоприменения
- Препринты сотрудников ИПП

Все материалы сотрудников ИПП Вы всегда можете найти на сайте **www.enforce.spb.ru**