Санкт-Петербургский государственный университет

**КОВАЛЕВ Роман Андреевич**

**Выпускная квалификационная работа**

**Автоматическое упрощение русскоязычных юридических текстов на основе извлечения фактической информации**

Уровень образования: бакалавриат

Направление 45.03.02 «Лингвистика»

Основная образовательная программа СВ.5106. «Прикладная, компьютерная и математическая лингвистика (английский язык)»

Научный руководитель:

к.ф.н., доц., Кафедра математической лингвистики,

Митрофанова Ольга Александровна

Рецензент:

доцент, к.ф.н., Кафедра общего языкознания имени Л.А. Вербицкой,

Блинова Ольга Владимировна

Санкт-Петербург

2023

Оглавление

[Введение 3](#_Toc136356835)

[Глава 1. Теоретические основания процедуры упрощения текстов 6](#_Toc136356836)

[1.1. Понятие семантической компрессии 6](#_Toc136356837)

[1.2. Подходы к адаптации текста 7](#_Toc136356838)

[1.3. Суммаризация текста как разновидность семантической компрессии 9](#_Toc136356839)

[1.4. Читабельность текста 14](#_Toc136356840)

[Глава 2. Подготовка лингвистических данных 16](#_Toc136356841)

[2.1. Анализ стилеметрических параметров текстов юридического дискурса 16](#_Toc136356842)

[2.2. Семантические и синтаксические характеристики адаптированных юридичеcких текстов 18](#_Toc136356843)

[2.3. Неоднозначность в юридических текстах 21](#_Toc136356844)

[2.4. Подготовка корпуса 23](#_Toc136356845)

[2.5. Лингвистическая разметка выбранных текстов 25](#_Toc136356846)

[2.6. Оценка сложности юридических текстов 27](#_Toc136356847)

[Глава 3. Эксперименты по упрощению юридических текстов 31](#_Toc136356848)

[3.1. Описание используемых алгоритмов 31](#_Toc136356849)

[3.2. Подготовка метрик качества упрощения текста 34](#_Toc136356850)

[3.3. Тестирование модели упрощения текста 39](#_Toc136356851)

[3.4. Экспертная оценка 42](#_Toc136356852)

[3.5. Анализ положительных и отрицательных результатов, обзор ошибок 45](#_Toc136356853)

[3.6. Анализ возможных ограничений упрощения юридических текстов 47](#_Toc136356854)

[Заключение 50](#_Toc136356855)

[Список литературы 51](#_Toc136356856)

[Список электронных ресурсов 54](#_Toc136356857)

# Введение

Упрощение текста (семантическая компрессия) – это процесс, включающий в себя лексико-семантическое и морфосинтаксическое упрощение исходного текста. Результат представляется в виде связного текста. Семантическая компрессия может осуществляться как на уровне слов (например, приведение от видового понятия к родовому, свертывание полной именной группы), так и на уровне предложений (экстрактивная и абстрактивная суммаризация текста).

Потребность в методах компрессии информации вызвана различием в скорости количественного увеличения словарного состава языка и скорости семантической деривации. Особый интерес для задач автоматического упрощения представляют тексты юридического дискурса. Упрощение особенно актуально для таких текстов, так как этот процесс обеспечивает понимание правовых отношений всеми участниками дискурса.

**Актуальность** исследования обусловлена повышенным интересом к методам семантической компрессии и представлениям юридического текста в более упрощенном и однозначном виде.

**Объект** исследования – корпус русскоязычных законов РФ на федеральном уровне с морфологической и синтаксической разметкой.

**Предмет** исследования – методы упрощения текста на лексико-семантическом и морфосинтаксическом уровне.

**Цель** исследования заключается в разработке и реализации алгоритма упрощения юридических текстов с целью повышения его удобочитаемости.

Для достижения цели исследования были сформулированы следующие **задачи**:

1. стилеметрический анализ современных юридических текстов;
2. отбор юридических текстов для включения в исследование;
3. разработка программного обеспечения для автоматического составления корпуса юридических текстов;
4. лингвистическая аннотация юридических текстов;
5. исследование и сравнение методов упрощения текста на лексико-семантическом и морфосинтаксическом уровнях;
6. разработка алгоритма для упрощения текста;
7. тестирование разработанного алгоритма;
8. исследование и применение метрик сложности текста применительно к адаптированным текстам на юридическую тематику
9. проведение оценки полученных результатов экспертами, анализ положительных и отрицательных результатов, анализ ошибок.

**Структура работы:**

Данная работа состоит из введения, трех глав, заключения и списка литературы. Во введении формулируется проблема, обосновывается актуальность исследования, отмечается практическая значимость работы, определяются цель, задачи и методы исследования, обозначается объект и предмет исследования.

**Первая глава** содержит теоретические основания исследования. В ней описывается понятие семантической компрессии, сравниваются различные подходы к адаптации текстов. Далее приводится и раскрывается понятие читабельности текста, а также понятие неоднозначности применительно к юридическим текстам.

**Вторая глава** посвящена подготовке лингвистических данных для проведения исследования. Представлены критерии для отбора юридических текстов для упрощения, приведено описание корпуса и его разметки. Также приводятся различные метрики, предназначенные для оценки сложности текста.

**Третья глава** посвящена практической части исследования, в ходе которой реализован алгоритм упрощения юридических текстов. В ней последовательно описывается и обосновывается каждый этап алгоритма. Также в этой главе представлены примеры работы алгоритма, оценка его работы по метрикам сложности, а также экспертная оценка читабельности полученных текстов. В конце главы представлен анализ положительных и отрицательных результатов работы, обзор ошибок, а также анализ возможных ограничений работы алгоритма.

В **заключении** обобщаются теоретические и практические результаты исследования и анализируются недостатки использованных методов.

# Глава 1. Теоретические основания процедуры упрощения текстов

## 1.1. Понятие семантической компрессии

Потребность в методах компрессии информации вызвана различием в скорости количественного увеличения словарного состава языка и скорости семантической деривации [Комарова 2005]. В процессе расширения знаний об окружающем мире инструменты языковой компрессии становятся все более актуальны.

В лингвистике распространены две трактовки компрессии. В широком понимании компрессия является одной из главных тенденций языка, непосредственно связанной с языковой экономией, так как ее результатом становится свернутая конструкция с сохранением объема передаваемой информации. В узком смысле компрессия – набор конкретных средств, служащих целям ее реализации. В этом смысле компрессия может осуществляться на всех языковых уровнях, в том числе семантическом, и на всех уровнях языка она представляет собой приближение к основному смыслу текста с помощью поверхностной информации о тексте [Kazantseva 2006].

На семантическом уровне языка компрессия связана с такими явлениями, как имплицитность, пресуппозиции и контекст [Шокина 2008]. Имплицитная информация может быть представлена в двух вариантах. Первый вариант объединяет случаи, когда содержательные элементы осознаются адресатом благодаря субъективным установкам и опыту человеческого сознания. Второй вариант имплицитности определяется сокращениями формальной структуры высказывания, при которых тот или иной знак редуцируется, но при этом его означаемое остается в зоне восприятия.

Одним из следствий семантической компрессии, как и других видов компрессии, является в той или иной степени потеря информации, содержащейся в исходном тексте [Ceglarek 2010]. Так как размерность значения в семантическом пространстве при компрессии всегда уменьшается, необходимо рассматривать семантическую компрессию с учетом качества созданного реферата в целом.

## 1.2. Подходы к адаптации текста

Одной из разновидностей семантической компрессии является адаптация текста. Адаптация текста – это процесс создания вторичного текста для читателей, которые по некоторым причинам не могут понять текст-источник [Первухина 2014]. Адаптированный текст, полученный в результате различных преобразований, содержит отсылку к той же предметной ситуации, что и исходный текст. При создании адаптированного текста его автор учитывает языковое сознание новой группы читателей.

Адаптированные тексты актуальны в тех специфических областях знаний, в которых доступность знаний затруднена. К таким сферам относятся:

* изучение иностранных языков;
* юриспруденция;
* медицинская сфера (например, адаптация текстов инструкций к лекарственным препаратам).

Стоит отметить, что при адаптации текста учитывается компетентность автора в сфере, к которой относится адаптируемый текст. Если автор не компетентен в данной области, то он предлагает интерпретацию текста, а не его адаптированный вариант.

Повышение понятности текста через его упрощение актуально для случаев, когда различия в фоновых знаниях специалистов в сфере и неспециалистов достигают значительной степени [Первухина 2013]. При упрощении текста передача информации осуществляется успешно, даже если невозможно восполнить знания адресата, который не является специалистом в данной сфере.

На лексико-семантическом и морфосинтаксическом уровнях упрощение текста может осуществляться следующими способами, которые различаются объектом преобразования (от лексической единицы до ситуации в целом):

* собственно упрощение, при котором требуется упрощение лексических единиц и стиля текста;
* разъяснение понятий – пояснение и уточнение специальных терминов;
* толкование ситуации – интерпретация, направленная на восполнение пробелов в фоновых знаниях адресата.

Большой вклад в развитие исследований по адаптации текста внесли достижения в корпусной лингвистике. Практика разработки и применения электронных корпусов продемонстрировала, что создание универсального корпуса, который обеспечил бы решение наиболее широкого круга задач, практически невозможно [Захаров, Богданова 2020].

Корпусная лингвистика оперирует как минимум тремя типами корпусов текстов:

* корпусы первого типа – эти корпусы отражают все многообразие речевой деятельности;
* корпусы второго типа – специфичные, отражают некоторые языковое или культурное явление в общественной речевой практике;
* корпусы третьего типа – специфичные, создаваемые для решения специальной задачи, например, для отладки систем машинного перевода.

Для задач адаптации текста наиболее важными оказываются так называемые корпусы третьего типа. Среди таких корпусов можно выделить параллельные корпусы – корпусы, в которых тексты в парах соотнесены по определенному критерию. В сопоставимых корпусах тексты объединены по некоторому признаку. Пары текстов «текст повышенной сложности – текст, более легкий для восприятия» в сопоставимых одноязычных корпусах позволяют выявить характеристики применительно к конкретному языку, свидетельствующие о сложности восприятия текста, а также указать на возможные пути его адаптации.

Так как в данной работе мы рассматриваем адаптацию с точки зрения повышения доступности знаний и легкости восприятия текста читателем, в дальнейшем мы используем термин «упрощение».

## 1.3. Суммаризация текста как разновидность семантической компрессии

Суммаризация текста, или автоматическое реферирование, как разновидность семантической компрессии представляет собой извлечение наиболее важных сведений из одного или нескольких документов [Батура, Бакиева 2019]. Входные данные в алгоритме автореферирования – это текст или несколько текстов, результатом алгоритма является сжатое представления этого текста – аннотация. Аннотация в то же время является краткой характеристикой этого документа.

К преимуществам автоматического реферирования текста относят:

1) облегчение поиска информации в крупных корпусах документов;

2) нейтральность и безоценочность аннотаций, сгенерированных автоматически;

3) упрощение индексирования документов;

4) возможность применения в классификации и кластеризации.

Решения задачи суммаризации текста можно подразделить на три направления: экстрактивные, абстрактивные и гибридные подходы [Батура, Бакиева 2019]. Один из основных критериев – глубина анализа текста. В экстрактивных подходах, которые также называются поверхностными, формируется «квази-реферат» – набор извлеченных из исходного текста наиболее информативных предложений. В абстрактивных подходах применяется глубинный анализ текста, при котором текст реферата генерируется на основе исходного с учетом морфологии, синтаксиса и семантики. В гибридных подходах используются как экстрактивные, так и абстрактивные методы суммаризации.

При экстрактивной суммаризации реферат полностью состоит из предложений, извлеченных из исходного текста. Значимость фрагментов, формирующих конечный реферат, может определяться по разным критериям, в том числе:

* по содержанию во фрагменте ключевых слов;
* по расположению фрагмента в исходном тексте;
* по наличию сигнальных фраз.

Пример исходного текста и реферата, составленного с помощью экстрактивной суммаризации, представлен ниже [Gusev 2020].

Исходный текст:

*На заседании Центризбиркома в четверг было принято постановление о разъяснении порядка работы международных наблюдателей на выборах депутатов Госдумы в декабре 2011-го и президента РФ в марте 2012-го. Постановление существенно ограничивает деятельность наблюдателей, которые и без того могут приехать в Россию только по приглашению ЦИК. «Иностранным наблюдателем не может быть, как указывается в нашей рекомендации, гражданин РФ», — цитирует РИА «Новости» члена ЦИК Елену Дубровину, выступившую на заседании комиссии в четверг. В этом документе функции наблюдателей также довольно подробно расписаны. Они получат право наблюдать за подсчетом голосов и знакомиться с бюллетенями. Кроме того, теперь они официально имеют право общаться не только с кандидатами и представителями партий, но с членами избирательных комиссий.*

Реферат:

*На заседании Центризбиркома в четверг было принято постановление о разъяснении порядка работы международных наблюдателей на выборах депутатов Госдумы в декабре 2011-го и президента РФ в марте 2012-го. Постановление существенно ограничивает деятельность наблюдателей, которые и без того могут приехать в Россию только по приглашению ЦИК.*

На первой стадии экстрактивной суммаризации исходная текстовая информация сводится к заданному числу фрагментов – единицам значения; ими могут быть категории, последовательности и темы. На второй стадии экстрактивного подхода определяются регулярные связи между единицами значения. Выводы и обобщения, составляющие структурную аннотацию текста, формируются на третьей стадии.

При использовании абстрактивных методов, в отличие от экстрактивных, на основе исходного текста генерируется новый, логически связанный текст. Этот способ также называется глубинным.

Ниже приведен пример исходного текста и реферата, составленного с помощью абстрактивной суммаризации [Gusev 2020].

Исходный текст:

*На заседании Центризбиркома в четверг было принято постановление о разъяснении порядка работы международных наблюдателей на выборах депутатов Госдумы в декабре 2011-го и президента РФ в марте 2012-го. Постановление существенно ограничивает деятельность наблюдателей, которые и без того могут приехать в Россию только по приглашению ЦИК. «Иностранным наблюдателем не может быть, как указывается в нашей рекомендации, гражданин РФ», — цитирует РИА «Новости» члена ЦИК Елену Дубровину , выступившую на заседании комиссии в четверг. В этом документе функции наблюдателей также довольно подробно расписаны. Они получат право наблюдать за подсчетом голосов и знакомиться с бюллетенями. Кроме того, теперь они официально имеют право общаться не только с кандидатами и представителями партий, но с членами избирательных комиссий.*

Реферат:

*Центризбирком ограничил работу международных наблюдателей на выборах депутатов Госдумы в декабре 2011-го и президента РФ в марте 2012-го. Постановление существенно ограничивает деятельность наблюдателей, которые и без того могут приехать в Россию только по приглашению ЦИК.*

Для реализации глубинного анализа рефератов необходимо наличие методов семантического или синтаксического разбора предложения. В первом случае формируются синтаксические деревья. Во втором случае также выполняется синтаксический разбор, но при этом порождения синтаксических деревьев не происходит; формируются семантические структуры, накапливающиеся в виде концептуальных графов в тезаурусах и базах знаний.

Благодаря многоуровневому структурированию текста можно подойти к задаче суммаризации текста разными путями. Абстрактивная суммаризация может предполагать как удаление и сокращение малозначащих смысловых единиц, так и уточнение реферата с помощью статистических методов.

Другая классификация методов суммаризации включает в себя следующие категории [Bharti 2017]:

* статистический подход;
* подход, основанный на машинном обучении;
* когерентный подход;
* графовый подход;
* алгебраический подход.

Статистический и когерентный подход используются для извлечения ключевых слов из текста и отношений согласованности между элементами текста. В алгебраическом способе суммаризации (например, в алгоритме LSA [Landauer 1998]) используются матрицы и собственные векторы. В графовом подходе элементы текста представляются в виде вершин графа, а семантические отношения между элементами текста обозначаются ребрами. В машинном обучении используется размеченный набор данных, а выходом алгоритма обучения является функция, которая аппроксимирует зависимость между элементами текста.

Несмотря на то, что суммаризация текста является разновидностью семантической компрессии, ее результатом не всегда становится упрощенный текст. Суммаризация решает задачу сокращения текста и выделения наиболее содержательный его частей, передающих основной смысл. Сложность восприятия текста читателем может остаться на том же уровне, если в ходе суммаризации лексико-семантическая и морфосинтаксическая сложность практически не изменились. Попытка использовать результаты автоматической суммаризации для упрощения текста была предпринята в работе [Sun 2023].

## 1.4. Читабельность текста

Задача упрощения текста тесно связано с оценкой его читабельности – легкости восприятия этого текста читателем. Традиционные метрики читабельности позволяют решить следующие задачи:

* соответствие читательской способности адресата;
* решение о необходимости упрощения текста или его отдельных частей.

В работе [Кисельников 2015] указывается на неоднозначность в трактовке термина «читабельность». В зарубежных работах, посвященных исследованию сложности текста, читабельность традиционно определяется как характеристика текста, определяемая на основе количественных параметров, например, количество знаков или слогов в словах. Однако читабельность можно также рассматривать как измерение того, насколько удобочитаемым является текст, на основании среднего уровня подготовки адресатов. В связи с этим большинство индексов читабельности используются в современных исследованиях для оценки сложности текстов не только для вычисления их количественных характеристик, но и уровня образования или возраста, необходимого для понимания текста целиком.

На сегодняшний день насчитывается более двухсот формул читабельности, адаптированных под особенности различных языков. Языковые особенности следует учитывать при автоматическом анализе читабельности текста.

Кроме количественных характеристик текста для оценки его читабельности также могут использоваться его словарные признаки, полученные с использование лексических минимумов. Такие признаки зачастую использоваться при адаптации учебных текстов. Как и индексы читабельности, словарные признаки, характеризующие сложность текста, могут быть адаптированы под различные языки.

Для русского языка широко используется градуальная серия лексических минимумов [Андрюшина 2015]. Эта система обеспечивает унифицированный контроль над овладением русским языком и объективную оценку уровня владения. Уровни владения русским языком включают в себя шесть ступеней: от элементарного уровня (ТЭУ) до четвертого сертификационного уровня (ТРКИ-4), а также опираются на ограниченный объем лексики, необходимый для освоения языка на определенном уровне. Таким образом, эта система в частности решает задачу упрощения текста, так как может быть использована в автоматическом упрощении текстов на лексическом уровне.

# Глава 2. Подготовка лингвистических данных

## 2.1. Анализ стилеметрических параметров текстов юридического дискурса

Юридические тексты представляют собой тексты официально-делового стиля, для которого характерны точность, обобщенность и стандартизированность изложения [Воронцова 2009]. В связи с этим в текстах официально-делового стиля семантическая структура максимально упрощена, чтобы устранить возможную неоднозначность в прочтении. Функции лиц и объектов в официально-деловых отношениях закреплены за определенными лексическими средствами, что упрощает построение логического представления юридического текста.

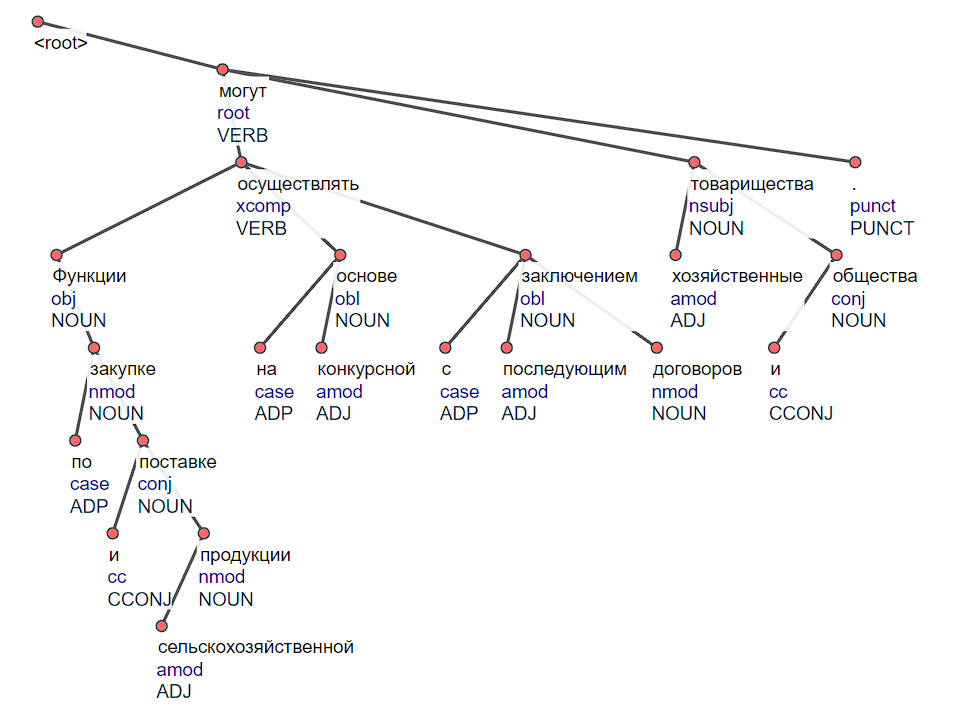
Стоит отметить, что стандартизированность изложения в юридических текстах приводит к тому, что они становятся менее читабельными – более сложными для восприятия читателем. Так как в задачи суммаризации входит упрощение текста, алгоритмы суммаризации должны увеличивать читабельность текста с учетом того, что все логические связи из исходного документа также отражены в реферате.

Большое количество современных работ по исследованию сложности текстов посвящены метрикам сложности правовых текстов [Блинова, Тарасов 2022]. В работе [Савельев 2018] продемонстрированы метрики сложности текста применительно к русскоязычным правовым актам. Как правило, лексическое разнообразие делает текст более сложным для восприятия. Однако в юридических текстах, несмотря на меньшее разнообразие и формальные повторы одних и тех же юридических терминов и наименований субъектов права, восприятие смысла текста усложнено. Таким образом, излишние повторы делают юридический текст менее читабельным.

Сложность структуры предложений также можно измерять с помощью расстояния между зависимыми словами. Построенные для юридических текстов деревья зависимостей имеют длинные «ветви», и зачастую слова в отношениях зависимости находятся в начале и конце предложения, что также затрудняет восприятие таких текстов читателем.

В приведенном ниже примере показано дерево зависимостей для предложения из текста юридической тематики (с применением синтаксического парсера UDPipe (https://lindat.mff.cuni.cz/services/udpipe/)), в котором есть зависимость между двумя словами, расположенными не рядом («могут» и «товарищества»):

*Функции по закупке и поставке сельскохозяйственной продукции могут осуществлять на конкурсной основе с последующим заключением договоров хозяйственные товарищества и общества.*



*Рис. 1. Дерево зависимостей для предложения из текста юридического стиля.*

Таким образом, стилистические черты, свойственные текстам юридического домена, указывают на повышенную синтаксическую и лексическую сложность. Эти проблемы необходимо учитывать при составлении упрощенного варианта текста, в особенности на этапе предобработки.

## 2.2. Семантические и синтаксические характеристики адаптированных юридичеcких текстов

Юридический дискурс непосредственно связан с нормализацией отношений между индивидами [Кожемякин 2010] и, как правило, регулирование этих отношений осуществляется с помощью различных текстов. Адаптация текста особенно актуальна для таких текстов, так как этот процесс обеспечивает понимание этих отношений всеми участниками дискурса.

Юридический текст может быть адаптирован на морфосинтаксическом и лексико-семантическом уровне. В рамках упрощения языка текста на лексико-семантическом уровне могут быть осуществлены следующие операции:

* использование более широко употребимой лексики вместо специальной терминологии ( «голосование» вместо «референдум»);
* замена абстрактных, неясно определенных понятий на их упрощенные варианты (замена понятие «частная собственность» на «собственность граждан»).

Синтаксические характеристики, свойственные адаптированным юридическим текстам, включают в себя более короткие предложения и упрощенные синтаксические конструкции.

Для упрощения юридических текстов также широко используется разъяснение абстрактных и специфических понятий. Такая модель упрощения часто встречается в текстах особого типа – юридических документах интерпретационного характера, или актах официального толкования права. Разъяснения используются в текстах для того, чтобы зафиксировать смысл коммуникативных действий участников дискурса и соответствующих правовых текстов.

Пример разъяснения понятия «референдум» можно обнаружить в Комментарии к Конституции РФ, где он раскрывается следующим образом:

*Народ, выражая свою волю на референдуме, свободных выборах, самостоятельно решает наиболее важные государственные и общественные вопросы, определяет состав избираемых им органов публичной власти, обеспечивает демократическую легитимацию власти в государстве.*

Более широкая операция – это толкование ситуации, при котором объясняются не только термины, но и целая ситуация, описанная в тексте. Такой вид упрощения не только восполняет недостаток знаний у адресата, но и предоставляет новую информацию о ситуации. Различие между разъяснением и толкованием заключается в объекте операции: в юридических текстах разъяснение – это объяснение терминов, в свою очередь толкование относится к ситуации в целом.

В адаптированном варианте текста могут быть применены все перечисленные выше виды адаптации, и, как следствие, признаки всех этих способов могут быть обнаружены в адаптированном тексте.

Стоит отметить, что в модели упрощении текста исходный текст обладает большей семантической плотностью. Напротив, в моделях разъяснения и толкования адаптированный текст обладает большей семантической плотностью, так как понятия и ситуации разворачиваются, и происходит экспликация денотативных сем терминов. Это демонстрирует следующий пример [Первухина 2013]:

Исходный текст:

*Земля и другие природные ресурсы могут находиться в частной, государственной, муниципальной и иных формах собственности.*

Адаптированный текст:

*Под частной собственностью в комментируемой статье понимается собственность граждан, их объединений, юридических лиц, не являющихся государственными и муниципальными организациями и предприятиями. Право частной собственности означает, что каждый вправе иметь имущество в собственности, владеть, пользоваться и распоряжаться им как единолично, так и совместно с другими лицами, и никто не может быть лишен своего имущества иначе как по решению суда, а принудительное отчуждение имущества для государственных нужд может быть произведено только при условии предварительного и равноценного возмещения.*

На лексическом уровне адаптированные юридические тексты обладают большим лексическим разнообразием. Понятие «лексическое разнообразие» в том числе обозначает, сколько повторов было допущено в конкретном тексте. К этим повторам приводят строгость формулировок и названий субъектов права. На лексическое разнообразие также может указывать количество уникальных слов в тексте. В этом заключается особенность правовых текстов от, например, учебных текстов. В художественных и учебных текстах увеличение лексического разнообразия ведет к повышению сложности текста [Савельев 2018].

На синтаксическом уровне на сложность текста также указывает максимальное расстояние между связанными членами предложения [Савельев 2018]. Структуру предложения можно представить в виде дерева, схема которого соответствует математическому графу, в котором вершины соответствуют словам, а ребра графа – отношениям подчинения между зависимыми элементами предложения. Путем таких построений это максимальное расстояние можно высчитывать автоматически. При адаптации юридических текстов наибольшее расстояние между зависимыми членами предложения сокращается в более значительной степени, чем при упрощении текстов, относящихся к другим сферам.

В данной работе применительно к текстам федеральных законов были использованы как методы упрощения на лексико-семантическом, так и морфосинтаксическом уровне. Лексико-семантическое и морфосинтаксическое упрощение происходит на этапе суммаризации; контекстные замены и нормализация именованных сущностей адаптируют текст на семантическом уровне.

## 2.3. Неоднозначность в юридических текстах

Языковая единица, высказывание или речевое произведение, обладающее более чем одним значением, считается неоднозначным. В работе [Блинова, Белов 2020] проводится различие между языковой и речевой неоднозначностью:

* языковая неоднозначность связана со способностью иметь разные смыслы, что является свойством языковых единиц;
* речевая неоднозначность – реализация этой потенциальной способности в конкретном высказывании.

Закрепление правовых норм и конкретных фактических обстоятельств в текстах письменных юридических документов требует от них ясности и определенности. Юридические тексты призваны разрешить двусмысленность, с которой специалисты в сфере права встречаются в практической деятельности. Однако неоднозначность как распространенное свойство языковых единиц приводит к появлению неоднозначности в текстах юридической тематики, и исследователи неоднозначности в текстах [Zeller 2018] связывают это с тем, что юристы используют в текстах естественный язык.

В юридических текстах неоднозначность наблюдается на нескольких уровнях. На лексико-семантическом уровне неоднозначность связана с наличием у слова или словосочетания нескольких связанных значений. Кроме случаев, когда языковая единица может быть неоднозначной в обыденном языке, выделяются также следующие случаи:

* языковая единица также имеет юридическую дефиницию, которая вводит различие между специальным пониманием термина и его значением в обыденном языке (например, термин «гражданский брак»);
* языковая единица имеет несколько существенно различающихся юридических дефиниций (определения термина «близкие родственники» в Уголовно-процессуальном кодексе РФ и Семейном кодексе РФ).

Кроме того, в юридических текстах неоднозначность может быть реализована на структурном (синтаксическом) уровне. Синтаксическая неоднозначность в примере ниже также может приводить к различным прочтениям текста [Блинова, Белов 2020]:

*Право собственности и иные вещные права на воздушные суда, морские суда, суда внутреннего плавания, космические объекты, подлежащие государственной регистрации, определяются по праву страны, где эти суда и объекты зарегистрированы.*

В этом случае «подлежащие государственной регистрации» может относиться как к словосочетанию «космические объекты», так и всей группе «воздушные суда, морские суда, суда внутреннего плавания, космические объекты».

Неоднозначность также может быть связана с возможностью установить кореферентность в тексте. Это понятие описывает отношение, при котором референциальные выражения ссылаются на один референт. Если отсутствует возможность однозначно установить в тексте кореферентность, то в нем наблюдается референциальная неоднозначность. Так в примере ниже форма «ему» может относиться к нескольким референтам:

*В случае если для подготовки ответа требуется более продолжительное время,* ***специалист администрации****, осуществляющий индивидуальное устное информирование, может предложить* ***заявителю*** *обратиться за необходимой информацией в письменном виде либо назначить другое удобное для* ***него*** *время для устного информирования.*

Таким образом, в текстах юридического стиля неоднозначность наблюдается на нескольких уровнях, и связана с тем, что эти тексты создаются с помощью естественного языка, единицам которого свойственна неоднозначность.

## 2.4. Подготовка корпуса

Современные корпусы юридических текстов предназначены для решения широкого круга задач. Среди корпусов по данной области знаний наиболее известны следующие: многоязычные корпусы Europarl (https://www.statmt.org/europarl/), United Nations Parallel Corpus (https://conferences.unite.un.org/UNCorpus), русскоязычные корпусы CorCodex, CorDeс, СorRIDA (https://www.plaindocument.org/corpora) [Блинова, Тарасов 2022; Савельев 2020]. В исследовании [Чернявская 2019] представлен параллельный корпус текстов для обучения студентов юридическому переводу. В работе [Митрофанова, Атугодаге 2023] проводится тематическое моделирование корпуса юридических текстов, в который были включены указы высших должностных лиц, постановления Верховного и Конституционного судов и законодательные акты Российской Федерации.

Возможность нескольких прочтений неотъемлема для федеральных законов РФ, так как их применение касается наиболее широкого круга случаев. Неопределенность таких текстов предоставляет пространство для интерпретации [Блинова 2020], и исследование их упрощения представляет особый практический интерес. В связи с этим в данном исследовании было принято решение об использовании текстов законодательных актов на федеральном уровне.

В качестве юридических текстов для корпуса были выбраны кодексы и федеральные нормативно-правовые акты РФ. На данный момент количество кодексов РФ насчитывает 21, однако в корпусе к ним отнесены 30 документов, так как в него включены и старые, и новые версии кодексов. Также некоторые из самых крупных кодексов в корпусе разделены на несколько частей. Кроме того, корпус включает в себя 132 федеральных конституционных закона и 6510 федеральных законов.

Материал корпуса включает в себя федеральные законы, принятые в период с 21 апреля 1991 года по 10 апреля 2018 года; в качестве метаданных для каждого текста представлена дата принятия и одобрения закона. Объем корпуса составил 14941511 токен.

Корпус представляет собой несбалансированный набор документов, так как количество кодексов и федеральных конституционных законов в десятки раз меньше, чем количество федеральных законов. Тем не менее, с точки зрения суммаризации текста и логики построения нормативно-правовых актов, задача суммаризации сводится к построению рефератов для глав внутри каждого кодекса и полных текстов федеральных законов, так как их структура идентична.

Разное количество типов документов приводит к тому, что определить оптимальную длину текстов в корпусе довольно сложно. Самый длинный из текстов в корпусе, Уголовно-процессуальный кодекс, в 800 раз крупнее самого короткого федерального закона. В связи с этим из корпуса были исключены тексты, основная часть которых состоит менее чем из 5 предложений.

## 2.5. Лингвистическая разметка выбранных текстов

Стандартизированность и строгая структура юридических текстов препятствует автоматической разметке, так как их формальная структура предполагает, что не вся информация подлежит суммаризации или семантической компрессии в целом. Каждый правовой акт включает в себя вступление и заключение с общей информацией о его принятии, к примеру:

*Арбитражный процессуальный кодекс*

*Российской Федерации*

*Принят Государственной Думой 14 июня 2002 года*

*Одобрен Советом Федерации 10 июля 2002 года*

…

*Президент Российской Федерации В.Путин*

*Москва, Кремль*

*24 июля 2002 года*

*N 95-ФЗ*

По этой причине было принято решение разметить в каждом акте три основные части: вступление с тегами <header> и </header>, основное содержание с тегами <body> и </body> и заключение с тегами <ending> и </ending>. Во время эксперимента упрощению подлежал только текст внутри тегов <body> и </body>. Также все разделы, подразделы, статьи, пункты и подпункты в основной части были размечены тегами <article\_name> и </article\_name>.

Корпусы, размеченные в формате UD (Universal Dependencies), в последнее время всё более активно используются при оценке морфосинтаксической сложности как при межъязыковом сопоставлении, так и при сравнении текстов (коллекций текстов) на одном языке. Поэтому в качестве базового инструмента разметки текстов в данной работе выбран UDPipe.

Лингвистическая разметка для каждого правового акта представлена в следующем формате:

*"text": "Президенту Российской Федерации и Правительству Российской Федерации привести свои нормативные правовые акты в соответствие с настоящим Федеральным законом."*

*"tokens": {*

*"token\_0": {*

*"word": "Президенту",*

*"lemma": "президент",*

*"pos": "NOUN",*

*"dep": "iobj",*

*},*

*"token\_1": {*

*"word": "Российской",*

*"lemma": "российский",*

*"pos": "ADJ",*

*"dep": "amod",*

*},*

*"token\_2": {*

*"word": "Федерации",*

*"lemma": "Федерация",*

*"pos": "PROPN",*

*"dep": "amod",*

*},*

*"token\_3": {*

*"word": "и",*

*"lemma": "и",*

*"pos": "CCONJ",*

*"dep": "conj",*

*},*

*"token\_4": {*

*"word": "Правительству",*

*"lemma": "Правительство",*

*"pos": "PROPN",*

*"dep": "conj",*

*},* …

## 2.6. Оценка сложности юридических текстов

Для оценки сложности юридических текстов в данной работе использованы несколько типов метрик, а именно:

* базовые метрики (средняя длина предложения, средняя длина словоформ и т. д.);
* учет слов разных частеречных классов;
* формулы читабельности;
* частотность лемм.

Данные метрики были отобраны с учетом стилистических особенностей текстов юридического дискурса, с помощью которых эти тексты выделяют в отдельный класс при выполнении задач классификации. Также при выборе метрик учтен опыт исследований сложности.

Базовые метрики, использованные в работе, включают в себя следующие показатели:

* доля словоформ, включающих 4 слога и более;
* средняя длина предложения в словах;
* средняя длина словоформы в слогах;
* отношение количества запятых к общему количеству токенов документа;
* доля гапаксов (лемм, встречающихся в тексте только один раз);
* доля лемм, встречающихся в тексте дважды.

В метрики читабельности входят адаптированная формула Флеша-Кинкейда (Flesch-Kincaid), [Flesch 1949] адаптированная формула SMOG (Simple Measure of Gobbledygook), [McLaughlin 1969] адаптированная формула подсчёта автоматизированного индекса читабельности ARI, индекс Дейла-Чейл (Dale-Chale) [Dale et al., 1948], индекс Колман-Лиау (Coleman-Liau) [Coleman et al., 1975]. Чем больше значение индекса читабельности, тем сложнее человеку воспринимать анализируемый текст.

Адаптированная формула Флеша-Кинкейда [Kincaid 1975] основана на традиционных метриках среднего количества слов в словах (ASL) и слов в предложении (ASW):

Индекс Дейла-Чейл, в свою очередь, опирается на сложность отдельных слов, представленных в тексте. Формула для вычисления индекса выглядит следующим образом:

Индекс Коулман-Лиау, основан на длине строковых значений в тексте. Формула для вычисления индекса выглядит следующим образом:

Здесь L обозначает среднее количество букв на каждые 100 слов, а S – среднее количество предложений на каждые 100 слов.

Индекс читабельности ARI также основан на отношении длин элементов предложения в тексте. Формула для вычисления индекса приведена ниже:

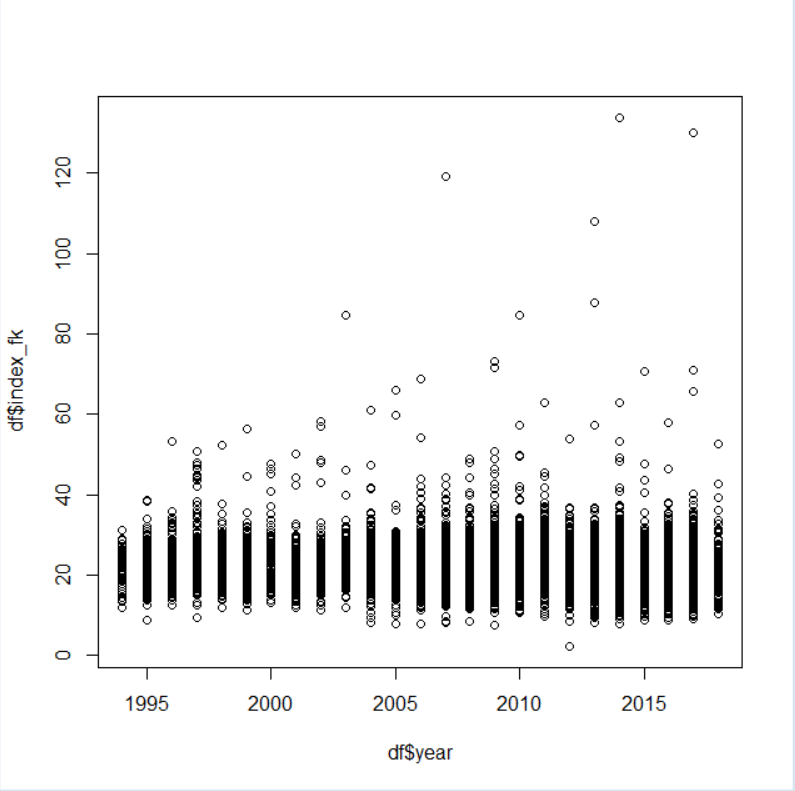
Наконец, адаптированная для русского формула SMOG (Simple Measure of Gobbledygook) выглядит следующим образом:

Остальные метрики сложности, использованные в работе, можно разделить на две большие группы: метрики лексической сложности и метрики морфосинтаксической сложности.

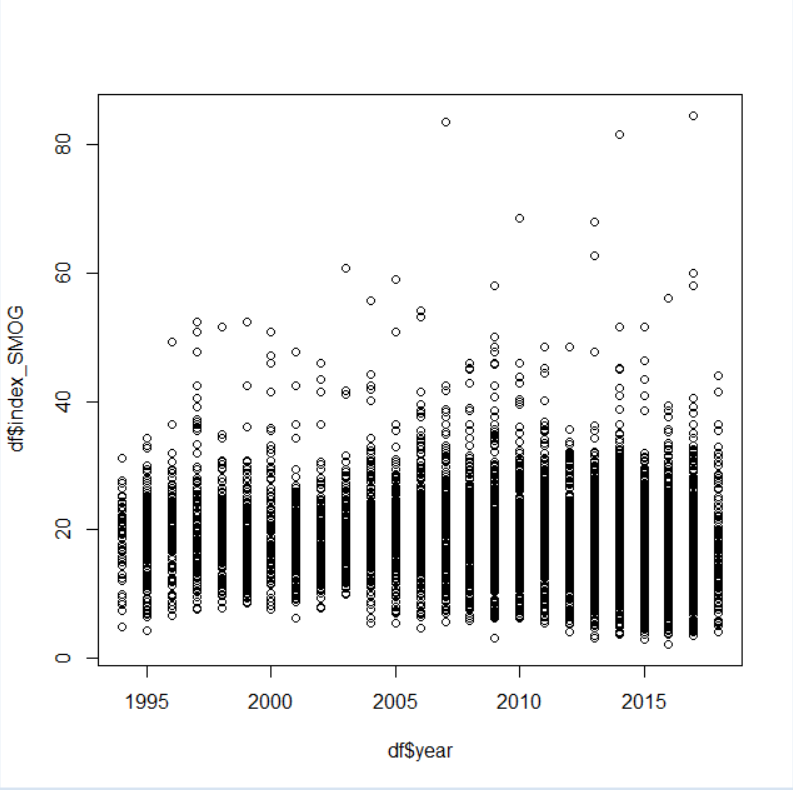
В качестве параметров оценки лексической сложности выбраны: значения простого TTR; значения метрик лексического разнообразия Yule's K и Yule's I; количество гапаксов.

Для оценки морфосинтаксической сложности выбраны, кроме прочего: количество неслужебных слов, в частности, существительных; значения соотношения количества существительных и глаголов; количество подчинительных и сочинительных союзов; количество тегов граммем родительного падежа, среднего рода, пассивного залога.

Данные для текстов из корпуса с оценкой читабельности по адаптированной формулам Флейша-Кинкейда и SMOG приведены ниже:



*Рис. 2. Оценка читабельности по Флейшу-Кинкейду для текстов разных лет.*

­­

*Рис. 3. Оценка читабельности по формуле SMOG для текстов разных лет.*

# Глава 3. Эксперименты по упрощению юридических текстов

## 3.1. Описание используемых алгоритмов

В данной работе была разработана гибридная методика упрощения текста. Эта методика опирается на суммаризацию как вид семантической компрессии. Стоит отметить, что суммаризация текста не всегда приводит к повышению его читабельности, так как целью суммаризации является исключение наименее содержательных частей текста. В связи с этим было принято решение о применении процедур обработки текста, позволяющих упростить суммаризированный текст. В качестве таких процедур были выбраны нормализация именованных сущностей и контекстные замены.

Таким образом, в данной работе будет рассмотрен алгоритм автоматического упрощения, который включает в себя следующие шаги:

1. суммаризация текста;
2. нормализация именованных сущностей;
3. использование контекстных замен.

Тем самым, предполагаемый нами подход к упрощению следует считать подходом, основанным на извлечении фактической информации.

На этапе суммаризации текста использована модель mBART [Liu et al., 2020]. mBART является модификацией архитектуры Трансформер. Текстовые данные на входе зашумляются и маскируются, а модель в свою очередь обучается восстанавливать тексты. Так как модель обучена на мультиязычном корпусе, ее параметры можно настраивать с учетом специфики каждого языка.

На следующем этапе осуществляется задача извлечения и связывания именованных сущностей. Именованная сущность – это слово или словосочетание, которое выделяет предмет или явление в ряде аналогичных предметов или явлений. Как правило, набор именованных сущностей почти совпадает с набором имен собственных в тексте. Именованные сущности в юридических текстах связаны со следующими объектами:

* названиями государств («Россия», «РФ», «Российская Федерация»);
* названиями организаций («Совет Федерации», «Государственная Дума»);
* личными именами («В. В. Путин», «Д. А. Медведев»).

На данный момент наиболее часто используемыми инструментами для работы с сущностями являются библиотека Natasha (метод NewsNERTagger) (https://natasha.github.io/), Stanza (https://stanfordnlp.github.io/stanza/) и Pullenti (https://www.pullenti.ru/). В данной работе для извлечения и нормализации именованных сущностей была использована библиотека Pullenti. К преимуществам этой библиотеки можно отнести ее многозадачность и простоту в установке и использовании. Кроме того, Pullenti предоставляет детальный разбор при выполнении задачи выявления именованных сущностей (Named Entity Recognition, NER) и извлекает вложенные сущности, предоставляя их в нормальной форме. На соревнованиях RuNormAS (https://www.dialog-21.ru/evaluation/2021/runormas/) и RuNNE (https://www.dialog-21.ru/evaluation/2022/runne/) по нормализации текстовых фрагментов и извлечению вложенных именованных сущностей в 2021 и 2022 году соответственно библиотека Pullenti заняла первые места.

Последний этап алгоритма – это использование контекстных замен. В данной работе было использовано решение об ограничении круга слов, для которых могут быть использованы контекстные замены. Первое ограничение накладывается на лексическую принадлежность слов к определенному домену. Так как в качестве объекта исследования выбраны юридические тексты, замены были осуществлены только для слов на юридическую тематику. Второе ограничение связано со сложностью слов. В данной работе для идентификации трудных слов был использован критерий количества слогов. Таким образом, все слова из 4 слогов и больше, которые относятся к юридической тематике, подлежали контекстным заменам.

Для осуществления замен используется дистрибутивно-семантическая модель ruscorpora\_upos\_cbow\_300\_20\_2019 из сервиса RusVectōrēs (https://rusvectores.org/ru). Удобство использования модели заключается в том, что оно может осуществляться без локального скачивания модели. API-сервис RusVectōrēs позволяет решить задачу контекстных замен в автоматическом режим. Для этого должны быть изначально заданы:

* идентификатор модели, в которую необходимо направить запрос;
* слово, для которого будет осуществляться замена;
* формат выходных данных (API-сервис поддерживает два формата выходных данных: csv с разделением через табуляцию и json).

API-сервис RusVectōrēs поддерживает следующие функции:

* получение списка семантически близких слов для заданного слова в заданной модели;
* вычисление коэффициента косинусной близости между парой слов в заданной модели.

Для решения задачи контекстных замен была использована первая функция API-сервиса. Из 10 семантически близких слов, которые предлагает модель, по возможности выбиралось слово из 3 слогов и меньше, также относящееся к юридической тематике.

Пример 10 семантически близких слов для прилагательного «нормативный»:

*{*

*'нормативно-правовой\_ADJ': '0.64',*

*'правовой\_ADJ': '0.619',*

*'подзаконный\_ADJ': '0.608',*

*'нормативно-технический\_ADJ': '0.596',*

*'соответствие\_NOUN': '0.589',*

*'норматив\_NOUN': '0.543',*

*'регламентировать\_VERB': '0.54',*

*'законодательство\_NOUN': '0.539',*

*'норма\_NOUN': '0.537',*

*'нормирование\_NOUN': '0.535'*

*}*

Прилагательное «правовой» на этом шаге алгоритма будет выбрано для контекстной замены, так как оно состоит из 3 слогов и не удовлетворяет критерию сложности слова.

## 3.2. Подготовка метрик качества упрощения текста

Для того чтобы количественно оценить, насколько проще стал текст в результате работы алгоритма, нужно было автоматически оценить сложность текста. Онлайн-сервисы, которые позволяют рассчитать базовые метрики сложности (средняя длина предложения, средняя длина словоформ), а также индексы читабельности, включают в себя:

* Текстометр (https://textometr.ru/);
* PolyLing (https://polyling.spbpu.com/);
* readability.io (https://github.com/ivbeg/readability.io).

Текстометр – это онлайн-сервис, позволяющий, среди прочего, определить уровень читабельности текста, а также уровень его лексического разнообразия [Лапошина, Лебедева 2021]. Инструмент соотносит текст с одним из уровней по шкале CEFR (https://www.coe.int/en/web/common-european-framework-reference-languages), основываясь на статистике по покрытию текста лексическими минимумами и частотными словами русского языка.

Тем не менее, к недостаткам Текстометра можно отнести то, что он не предоставляет достаточной информации о читабельности текста: инструмент не возвращает информацию даже по базовым метрикам (средняя длина слова, средняя длина предложения). Кроме того, использование Текстометра на практике усложняется тем, что к нему нет возможности обратиться посредством API-запросов, что значительно упростило бы локальные эксперименты.

Сервис PolyLing – это интерактивный цифровой инструмент, который предназначен для повышения качества текстов [Никулина, Черкас, Козина, Бойко, Дмитриева 2023]. Под качеством текста подразумеваются следующие характеристики:

* лексическое разнообразие (лексико-грамматический состав предложений, метрики морфологической, лексической и синтаксической сложности и т. д.);
* визуальное восприятие (расположение иллюстрационных материалов, оптимальный размер и тип шрифта и межстрочных интервалов и т. д.).

Результатом работы инструмента является усредненное значение метрик лексического разнообразия и визуального восприятия, а также краткая рекомендация по улучшению восприятия текста читателем. Однако на выходе инструмент возвращает ограниченную количественную информацию о лексической и синтаксической сложности текста, которая представляет собой усредненное значение оценок по нескольким метрикам читабельности.

Сервис readability.io предназначен для оценки читабельности текстов, которая позволяет автоматически определить сложность текстов. Помимо пяти алгоритмов оценки читабельности текстов (адаптированная формула Флеша-Кинкейда адаптированная формула SMOG, адаптированная формула подсчёта автоматизированного индекса читабельности ARI, индекс Дейла-Чейл, индекс Колман-Лиау), сервис также возвращает количественные значения относительно других метрик, например, среднее число слов и слогов на предложение.

Алгоритм сервиса readability.io был обучен на следующих группах текстов, для каждого из которых был размечен необходимый для понимания уровень образования. Алгоритм был обучен для русскоязычных текстов; каждая из формул была подобрана с учетом особенностей текстов, написанных на русском языке. Эти группы текстов включают в себя:

* тексты для внеклассного чтения;
* размеченные художественные тексты;
* тексты законов.

Результатом анализа является следующий набор характеристик, представленный в формате json:

*{*

*'metrics': {*

*'wsyllabes': {*

*'1': 8,*

*'3': 17,*

*'2': 21,*

*'5': 15,*

*'4': 7,*

*'6': 4*

*},*

*'c\_share': 26.38888888888889,*

*'chars': 617,*

*'avg\_slen': 36.0,*

*'spaces': 79,*

*'n\_syllabes': 228,*

*'n\_words': 72,*

*'letters': 499,*

*'n\_sentences': 2,*

*'n\_complex\_words': 19,*

*'n\_simple\_words': 53,*

*'avg\_syl': 3.1666666666666665*

*},*

*'status': 0,*

*'lang': 'ru',*

*'debug': '0',*

*'indexes': {*

*'grade\_SMOG': 'Аспирантура, второе высшее образование, phD',*

*'grade\_ari': 'Аспирантура, второе высшее образование, phD',*

*'index\_fk': 25.914666666666662,*

*'grade\_cl': '4 - 6 курсы ВУЗа (возраст примерно: 20-22 лет)',*

*'grade\_fk': 'Аспирантура, второе высшее образование, phD',*

*'index\_cl': 16.815833333333337,*

*'grade\_dc': 'Аспирантура, второе высшее образование, phD',*

*'index\_dc': 24.39466666666667,*

*'index\_ari': 22.443277777777773,*

*'index\_SMOG': 27.30026605374707*

*}*

*}*

В первой части возвращаемых данных («metrics») представлены базовые количественные метрики сложности. Метрика n\_complex\_words высчитывает количество сложных для восприятия слов в тексте по критерию количества слогов: все слова из четырех слогов и более определяются как трудные для восприятия; c\_share определяется как отношение n\_complex\_words к n\_words (общему количеству слов в анализируемом тексте).

Во второй части («indexes») представлены значения пяти индексов читабельности. Параметры с index в названии – это количественные значения читабельности согласно каждому индексу, и они принимают значения от 1 до бесконечности. Параметры, в названии которых присутствует grade, отражают уровень образования, необходимый для понимания текста.

К преимуществам сервиса readability.io, отличающим его от других инструментов, можно отнести следующие:

* алгоритм был обучен, в частности, на текстах юридического домена, что делает его более подходящим для задачи определения сложности текстов в данной работе;
* предоставленный для свободного использования API-сервис позволяет ускорить автоматическую обработку текстов и совершать ее локально;
* API-сервис возвращает подробную информацию как о количественных характеристиках текста (средняя длина предложения, количество слов), так и количественные значения индексов читабельности.

Поэтому в данной работе для оценки исходных текстов и их упрощенных вариантов было принято решение использовать именно этот сервис.

## 3.3. Тестирование модели упрощения текста

На первом шаге эксперимента была предпринята суммаризация для полных юридических текстов. На вход модели mBART подавался целый текст в пределах тегов <body> и </body>.

В ходе этой части эксперимента были получены следующие результаты:

Таблица 1. Значения индексов читабельности для исходных текстов и рефератов, полученных в результате суммаризации текста в целом.

|  |  |  |
| --- | --- | --- |
| **Индекс читабельности** | **Исходные тексты** | **Рефераты** |
| Index\_fk | 20,31 | 21,39 |
| Index\_cl | 17,44 | 16,21 |
| Index\_dc | 19,51 | 20,51 |
| index\_ari | 18,64 | 18,84 |
| Index\_SMOG | 16,91 | 21,71 |

По четырем из пяти индексов значение читабельности суммаризированных текстов оказалось больше, чем то же значение для исходных текстов. В связи с этим было принято решение решить суммаризировать не целые тексты в пределах тегов <body> и </body>, а отдельные части текстов, заключенные в теги <article\_name> и </article\_name>. Затем рефераты, полученные для каждой отдельной части, снова объединялись в полный текст, для которого осуществлялись нормализация именованных сущностей и контекстные замены.

На этом шаге были получены следующие результаты по индексам читабельности:

Таблица 2. Значения индексов читабельности для исходных текстов и рефератов, полученных в результате суммаризации отдельных частей текста.

|  |  |  |
| --- | --- | --- |
| **Индекс читабельности** | **Исходные тексты** | **Рефераты** |
| index\_fk | 20,31 | 19,38 |
| index\_cl | 17,44 | 15,76 |
| index\_dc | 19,51 | 17,35 |
| index\_ari | 18,64 | 15,9 |
| index\_SMOG | 16,91 | 16,36 |

Таким образом, для повышения читабельности юридического текста необходимо использовать суммаризацию для отдельных частей текста, соответствующих разделам, статьям, подстатьям и т. д.

В ходе этой части эксперимента также выяснилось, что mBART осуществлял суммаризацию преимущественно для статей и подстатей ближе к началу текста. Статьи и подстатьи в конце закона инструмент возвращал в том же виде. Вероятно, это связано с тем, что статьи и подстатьи в конце федерального закона короче, чем первые статьи, и содержат меньше информации, которую можно суммаризировать.

Следующий шаг эксперимента включал в себя нормализацию именованных сущностей и контекстные замены. Пример работы алгоритма на этом шаге представлен ниже.

Исходный текст:

*Статья 2. Президенту Российской Федерации и Правительству Российской Федерации привести свои нормативные правовые акты в соответствие с настоящим Федеральным законом.*

Упрощенный вариант:

*Статья 2. Президенту РФ и Правительству РФ привести свои правовые правовые акты в соответствие с данным Федеральным законом.*

В этом тексте были определены следующие именованные сущности: ПРЕЗИДЕНТ (тип − PERSON), ПРАВИТЕЛЬСТВО (тип − ORGANIZATION), РОССИЙСКАЯ ФЕДЕРАЦИЯ (тип − GEO). Нормализация была осуществлена для именованной сущности РОССИЙСКАЯ ФЕДЕРАЦИЯ и приведена к нормальной форме РФ.

Для контекстных замен алгоритмом были выбраны слова «нормативные», «правовые» и «настоящим». Каждое из этих слов соответствует двум заранее выбранным критериям:

* сложность слова (4 слога и более);
* связь с юридическим доменом (слова «нормативный» и «правовой» непосредственно связаны с юридической тематикой, а слово «настоящий» часто встречается в клише, используемых в юридических текстах).

Инструмент RusVectōrēs для определения семантически близких слов определил замену для слова «нормативный» на «правовой», а для слова «настоящий» - на «данный». Что касается слова «правовой» в исходном тексте, среди десяти семантически близких слов не обнаружилось слова, которое удовлетворяло бы двум критериям, представленным выше. В связи с этим в упрощенном варианте текста слово «правовой» встретилось два раза подряд.

По результатам эксперимента были получены следующие результаты для индексов читабельности:

Таблица 3. Значения индексов читабельности для исходных текстов и упрощенных вариантов.

|  |  |  |
| --- | --- | --- |
| **Индекс читабельности** | **Исходные тексты** | **Упрощенные варианты** |
| index\_fk | 20,31 | 19,16 |
| index\_cl | 17,44 | 14,73 |
| index\_dc | 19,51 | 17,2 |
| index\_ari | 18,64 | 12,8 |
| index\_SMOG | 16,91 | 15,34 |

Таким образом, в ходе эксперимента были получены упрощенные варианты для всех текстов из корпуса. Для каждого из них были осуществлены суммаризация (для отдельных частей текста), нормализация именованных сущностей и контекстные замены. Как показали метрики читабельности, примененные к исходным текстам и их упрощенным вариантам, юридические тексты стали проще для восприятия.

## 3.4. Экспертная оценка

Для того чтобы провести экспертную оценку полученных результатов упрощения, были выработаны следующие критерии:

* читабельность текста – насколько просто воспринимается текст при чтении (от 1 до 5, где 1 – воспринимается тяжело, а 5 – легко читается);
* неоднозначность текста – вероятность неясности при прочтении юридического текста (от 1 до 5, где 1 – возможность многих трактовок исключена, а 5 – текст очень неоднозначен).

Для оценки качества упрощения были выбраны 10 текстов из корпуса и их упрощенные варианты. Чтобы облегчить оценку для экспертов, выбор был сосредоточен на текстах длиной от 5 до 20 предложений в основной части (между тегами <body> и </body>).

В качестве экспертов выступили трое сотрудников Юридической клиники СПбГУ. Эксперты должны были оценить тексты по указанным выше критериям и дать краткий комментарий о возможности применения упрощенных вариантов законов в практической деятельности.

В ходе экспертной оценки были получены следующие значения читабельности для 10 текстов, в таблице ниже указаны усредненные оценки от экспертов:

Таблица 4. Экспертная оценка читабельности исходных текстов и их упрощенных вариантов.

|  |  |  |
| --- | --- | --- |
| **Номер текста** | **Читабельность исходного текста** | **Читабельность упрощенного текста** |
| **Среднее значение** | **4,33** | **4,67** |
| 1 | 4,67 | 5 |
| 2 | 4,33 | 4,33 |
| 3 | 4 | 4 |
| 4 | 4,33 | 5 |
| 5 | 4,33 | 4,67 |
| 6 | 3,67 | 4,33 |
| 7 | 4,33 | 5 |
| 8 | 5 | 5 |
| 9 | 4,33 | 4,67 |
| 10 | 3,67 | 4,67 |

Следует отметить, что читабельность двух текстов из десяти была оценена так же, как и читабельность их упрощенных вариантов. Большинство текстов, по мнению экспертов, в результате эксперимента стали проще для восприятия.

Для неоднозначности в представленных текстах были получены следующие оценки:

Таблица 5. Экспертная оценка неоднозначности исходных текстов и их упрощенных вариантов.

|  |  |  |
| --- | --- | --- |
| **Номер текста** | **Неоднозначность исходного текста** | **Неоднозначность упрощенного текста** |
| **Среднее значение** | **1,67** | **4** |
| 1 | 2,33 | 4,33 |
| 2 | 1,33 | 4 |
| 3 | 2,33 | 3,33 |
| 4 | 1,33 | 3,33 |
| 5 | 2 | 4 |
| 6 | 2 | 4,67 |
| 7 | 2,67 | 5 |
| 8 | 1,33 | 4 |
| 9 | 1 | 3,67 |
| 10 | 1,67 | 4,67 |

Как следует из оценок, эксперты в основном посчитали исходные тексты однозначными, при этом неоднозначность полученных в ходе эксперимента упрощенных текстов получила более высокую оценку.

Таким образом, экспертная оценка показала, что юридические тексты в ходе эксперимента стали легче для восприятия, но их неоднозначность возросла в значительно большей степени.

## 3.5. Анализ положительных и отрицательных результатов, обзор ошибок

Тестирование модели на корпусе текстов федеральных законов и кодексов РФ показало, что данный алгоритм может применяться для упрощения текстов юридического домена. Юридические тексты, к которым были применены суммаризация текста, а затем нормализация именованных сущностей и контекстные замены, по всем использованным метрикам сложности и индексам читабельности показали большую удобочитаемость, чем соответствующие им исходные тексты.

Тем не менее, результаты некоторых шагов алгоритма могут повлиять на логическую структуру адаптированного текста. Так, например, не любое ближайшее слово, которое возвращает для сложного слова векторная модель, может использоваться в качестве полноценного синонима. Распространенная ошибка в работе модели – замена слова на его гипоним, который не соответствует описываемой ситуации:

* *правообладатель\_NOUN – патентообладатель\_NOUN*

Для некоторых слов векторная модель возвращала в качестве семантически ближайшего слова его антоним:

* *безвозмездный\_ADJ* – *возмездный\_ADJ*

Вероятно, эти ошибки в работе модели можно было бы избежать следующими способами:

* осуществление контекстных замен не для одного токена, а нескольких слотов, идущих подряд;
* ограничение на антонимы среди семантически ближайших слов.

На этапе суммаризации модель mBART возвращала для некоторых разделов текст в неизменном виде. Так как на следующих после суммаризации этапах синтаксических преобразований не осуществлялось, в таком случае на синтаксическом уровне текст не подвергался никаким изменениям, даже если по синтаксическим характеристикам этот текст можно было упростить.

Экспертная оценка продемонстрировала, что при использовании алгоритма значительно возрастает неоднозначность текста. Увеличение количества трактовок ведет к ограничению возможностей применения адаптированных текстов в практической деятельности.

## 3.6. Анализ возможных ограничений упрощения юридических текстов

Возможные ограничения упрощения текстов юридического домена связаны с особенностью их структуры и применения на практике.

Одной из важных черт текстов юридического домена является стандартизированность изложения, которая приводит к повышенной синтаксической сложности. Модели BART, и mBART в частности, обученные на текстах относящихся не только к юридической сфере, хуже справляются с суммаризацией текстов на юридическую тематику. Кроме того, наличие терминологии и специфических стандартизированных выражений делает эти тексты особенно сложными при решении задачи суммаризации автоматическим способом.

Также стоит отметить, что эксперты в юридической сфере, как правило, не считают тексты федеральных законов неоднозначными. Если норма вызывает разночтения судами разных инстанций или органов, то эти суды выпускают постановления и обзоры судебной практики, в которых предоставляют собственное толкование этой нормы. Из этого следует, что специалисты юридической сферы не только не считают большинство текстов законов неоднозначными, но и ликвидируют разночтения при их наличии посредством дополнительных документов, таких, как обзоры судебной практики. Таким образом, на уровне федеральных законов юристы не ожидают от текстов того, что они сами по себе будут давать возможность для большого числа трактовок.

Как следует из комментариев от экспертов Юридической клиники СПбГУ, разные статьи в рамках одного кодекса указывают на разный состав правонарушений. В качестве примера можно привести следующую статью из Уголовного кодекса РФ, в первом и втором пункте которой приведены описания преступлений, совершенные в разных условиях:

*УК РФ, Статья 139. Нарушение неприкосновенности жилища*

*1. Незаконное проникновение в жилище, совершенное против воли проживающего в нем лица, - наказывается штрафом в размере до сорока тысяч рублей или в размере заработной платы или иного дохода осужденного за период до трех месяцев, либо обязательными работами на срок до трехсот шестидесяти часов, либо исправительными работами на срок до одного года, либо арестом на срок до трех месяцев.*

*2. То же деяние, совершенное с применением насилия или с угрозой его применения, - наказывается штрафом в размере до двухсот тысяч рублей или в размере заработной платы или иного дохода осужденного за период до восемнадцати месяцев, либо исправительными работами на срок до двух лет, либо принудительными работами на срок до двух лет, либо лишением свободы на тот же срок.*

Этот принцип описания возможных правонарушений был учтен в данной работе, так как операции с элементами текста больше, чем слово, осуществлялись в рамках раздела, статьи, пункта и так далее. Что касается описания разных составов правонарушения внутри одного неделимого пункта или статьи, определение самых важных частей текста, которые должны остаться в законе при его упрощении, должно опираться на более глубокий семантический анализ. Как следствие, учет структуры текста при автоматическом упрощении не гарантирует, что все особенности правоприменения будут учтены в конечном варианте текста.

Несмотря на то что федеральный закон посвящен одной, очень узкой тематике, в его пределах подробно расписаны различные случаи правоприменения. Однако в практической деятельности для юристов нет необходимости использовать все разделы или статьи закона для его применения. Таким образом, специалисты юридической сферы менее заинтересованы в упрощении юридических текстов, так как при их прочтении учитывается не целый текст, а только его части, относящиеся к конкретному случаю правоприменения.

Тем не менее, эксперты отмечают, что применение упрощенных вариантов юридических текстов допустимо при относительно несложных правовых моментах, например:

* заключении договоров об оказании услуг;
* подписании договоров аренды;
* составление документов о подрядах или поставках.

Эти типы документов объединяет то, что сторонами в таких договорах являются физические лица. С точки зрения согласования договора между двумя физическими лицами или физическим лицом и юридическим, особенно важно полное понимание сторонами своих прав и обязанностей. Для неспециалиста в юридических вопросах будет особенно важно по меньшей мере понимать суть договоренности; в таком случае легкость восприятия текста и отсутствие специальных юридических терминов оказывается наиболее значимыми характеристиками текста.

# Заключение

В рамках данного исследования было проведен стилеметрический современных русскоязычных юридических текстов. Были отобраны юридические тексты для включения в исследования, а также разработано программное обеспечение для автоматического составления корпуса. Для текстов в корпусе была проведена лингвистическая аннотация. Было проведено исследование и сравнение методов синтаксического и семантического упрощения текста. В ходе исследования был разработан алгоритм упрощения текста, основанный на абстрактивной суммаризации, нормализации именованных сущностей и использовании контекстных замен. Были выбраны метрики автоматической оценки сложности текста. Данный алгоритм был протестирован на корпусе юридических текстов, для которых затем была проведена автоматическая оценка их сложности. Была проведена экспертная оценка, анализ положительных и отрицательных результатов, обзор ошибок. По результатам исследования разработанный алгоритм упрощения продемонстрировал возможность его применения для упрощения текстов на юридическую тематику и повышения их читабельности.

В дальнейших исследованиях предполагается обучение модели для суммаризации текста на текстах юридической тематики, а также применение алгоритма упрощения для юридически текстов, регулирующих правовые отношения, одна или обе стороны в которых являются физическими лицами.

Таким образом, разработанный в ходе работы алгоритм можно использовать для упрощения юридических текстов и повышения их читабельности.

# Список литературы

1. Андрюшина Н. П., Козлова Т. В. Лексический минимум по русскому языку как иностранному. Элементарный уровень. Общее владение. – СПб.: Златоуст. – 2015.
2. Батура Т. В., Бакиева А. М. Методы и системы автоматического реферирования текстов // Новосибирск: ИПЦ НГУ. – 2019.
3. Блинова О. В., Белов С. А. Языковая неоднозначность и неопределенность в русских правовых текстах. // Вестник Санкт-Петербургского университета. Право. – 2020. – Т. 11. – №. 4. – С. 774-812.
4. Блинова О. В., Тарасов Н. А. Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности // Компьютерная лингвистика и интеллектуальные технологии. – 2022. – С. 1017-1028.
5. Воронцова Т. А. Элементарная стилистика. // Ижевск: УдГУ. – 2009.
6. Захаров В. П., Богданова. Корпусная лингвистика: учебник. 3-е изд., перераб. // СПб.: Изд-во С.Петерб. Ун-та. – 2020.
7. Кисельников А. С. К проблеме характеристик текста: читабельность, понятность, сложность, трудность. // Тамбов: Грамота. – 2015. – Т. 53. – №11. – С. 79-84.
8. Кожемякин Е. А. Юридический дискурс как культурный феномен: структура и смыслообразование // URL: http://konference.siberia-expert.com/publ/doklad\_s\_obsuzhdeniem\_na\_sajte/kozhemjakin\_e\_a\_juridicheskij\_diskurs\_kak\_kulturnyj\_fenomen\_struktura\_i\_smysloobrazovanie/2-1-0-28 (дата обращения: 17.05.2023)
9. Комарова С. С. Семантическая компрессия в прагматике высказывания: на материале немецкоязычной прессы : дис. – Самара: автореф. дис.… канд. филол. наук, 2005.
10. Лапошина А. Н., Лебедева М. Ю. Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. – 2021. – Т. 19. – №3. – С. 330-345.
11. Митрофанова О. А., Атугодаге М. М. Динамическое тематическое моделирование русскоязычного корпуса юридических документов // Terra Linguistica. – 2023. – Т. 14. – №. 1: Инженерно-лингвистические технологии в исследованиях текста.
12. Никулина Е.Р., Черкас А. В., Козина Е. Д., Бойко А. В., Дмитриева Л. А. Разработка сервиса для оценки удобочитаемости текста с применением технологий машинного обучения // URL: https://cyberleninka.ru/article/n/razrabotka-servisa-dlya-otsenki-udobochitaemosti-teksta-s-primeneniem-tehnologiy-mashinnogo-obucheniya (дата обращения: 23.05.2023)
13. Первухина С. В. Виды адаптации текста // Вестник ЮУрГУ. Серия «Лингвистика». – 2014. – Т. 11. – №1. – С. 97-100.
14. Первухина С. В. Семантические и синтаксические характеристики адаптированного текста в юридическом дискурсе. // Тамбов: Грамота. – 2013. – Т. 22. – №4. – С. 136-140.
15. Савельев Д. А. Исследование сложности предложений, составляющих тексты правовых актов органов власти Российской Федерации // Право. Журнал Высшей школы экономики. – 2020. – № 1. – С. 50–74.
16. Савельев Д. А., Кучаков Р. К. Сложность правовых актов в России. – 2018.
17. Чернявская О. Г. Параллельный корпус текстов при обучении юридическому переводу // Конвергентные технологии ХХI: вариативность, комбинаторика, коммуникация. – 2019. – С. 212-215.
18. Шокина А. Б. Языковая компрессия в рекламном тексте // Медиаскоп. – 2008. – Т.10. – №.2. – С. 8.
19. Bharti S. K., Babu K. S. Automatic keyword extraction for text summarization: A survey // arXiv preprint arXiv:1704.03242. – 2017.
20. Ceglarek D., Haniewicz K., Rutkowski W. Semantic compression for specialised information retrieval systems //Advances in Intelligent Information and Database Systems. – 2010. – С. 111-121.
21. Coleman M., Liau, T. L. A computer readability formula designed for machine scoring. // Journal of Applied Psychology – 1975. – Т. 60. – С. 283–284.
22. Dale E., Chall J. A Formula for Predicting Readability. // Educational Research Bulletin. – 1948. – Т. 27. – №. 1. – С. 11-20.
23. Flesch R. The art of readable writing. // NY: Harper & Row. – 1949.
24. Gusev I. Dataset for automatic summarization of Russian news // Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Proceedings 9. – Springer International Publishing. – 2020. – С. 122-134.
25. Kazantseva A. An approach to summarizing short stories //Student Research Workshop. – 2006. – С. 55-62.
26. Kincaid J. P. et al. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. // Institute for Simulation and Training. – 1975.
27. Landauer T. K., Foltz P. W., Laham D. An introduction to latent semantic analysis // Discourse processes. – 1998. – Т. 25. – №. 2-3. – С. 259-284.
28. Liu Y. et al. Multilingual denoising pre-training for neural machine translation //Transactions of the Association for Computational Linguistics. – 2020. – Т. 8. – С. 726-742.
29. McLaughlin G. H. SMOG grading-a new readability formula // Journal of reading. – 1969. – Т. 12. – №. 8. – С. 639-646.
30. Sun R., Yang Z., Wan X. Exploiting Summarization Data to Help Text Simplification // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. – 2023. – С. 39-51.
31. Zeller T. Detecting ambiguity in statutory texts. BSc Thesis. 2018.

# Список электронных ресурсов

1. Проект Natasha – набор Python-библиотек для обработки текстов на естественном русском языке. // URL: https://natasha.github.io/ (дата обращения: 20.04.2023)
2. Соревнование RuNNE. // URL: https://www.dialog-21.ru/evaluation/2022/runne/ (дата обращения: 20.04.2023)
3. Соревнование RuNormAS. // URL: https://www.dialog-21.ru/evaluation/2021/runormas/ (дата обращения: 20.04.2023)
4. Текстометр – анализ сложности текста. // URL: https://textometr.ru/ (дата обращения: 22.05.2023)
5. Шкала CEFR. // URL: https://www.coe.int/en/web/common-european-framework-reference-languages (дата обращения: 26.05.2023)
6. CorCodex, CorDec, CorRIDA. // URL: https://www.plaindocument.org/corpora (дата обращения: 22.05.2023)
7. Europarl. // URL: https://www.statmt.org/europarl/ (дата обращения: 22.05.2023)
8. PolyLing. // URL: https://polyling.spbpu.com/ (дата обращения: 12.03.2023)
9. Pullenti. // URL: https://www.pullenti.ru/ (дата обращения: 20.04.2023)
10. Readability.io. // URL: https://github.com/ivbeg/readability.io (дата обращения: 28.05.2023)
11. RusVectōrēs: семантические модели для русского языка. // URL: https://rusvectores.org/ru (дата обращения: 22.05.2023).
12. Stanza. // URL: https://stanfordnlp.github.io/stanza/ (дата обращения: 20.04.2023)
13. UDPipe. // URL: https://lindat.mff.cuni.cz/services/udpipe/ (дата обращения: 12.03.2023)
14. United Nations Parallel Corpus. // URL: https://conferences.unite.un.org/UNCorpus (дата обращения: 22.05.2023)