Корпус русских локальных документов и актов CorRIDA: цели формирования, состав, структура

С.А. Белов[1], О.В. Блинова[1], В.Б. Гулида[1], В.И. Зубов[1],
Е.Ю. Ларионова[2], П.С. Толстикова[3]

[1] Санкт-Петербургский государственный университет,
[2] Европейский Университет в Санкт-Петербурге,
[3] Институт лингвистических исследований РАН

s.a.belov@spbu.ru, o.blinova@spbu.ru, v-gulida@yandex.ru,
vladzubov21@gmail.com, e.j.larionova@gmail.com,
hokori.chan@gmail.com

Аннотация

В статье описывается начальный этап создания лингвистически размеченного корпуса русских локальных документов и актов CorRIDA. В повседневной жизни носители русского языка всё чаще сталкиваются с необходимостью читать и подписывать различные официальные документы. Обычно это так называемые локальные документы, например, «Договоры на оказание платных услуг» или «Информированные согласия». Однако язык локальных документов исследован недостаточно и практически не рассматривался с применением корпусных методов. Существующие корпусы русского языка пока не предоставляют возможностей для систематического анализа языка документа. Это связано в том числе с проблемами жанровой классификации и разметки нехудожественных текстов. Поэтому формирование корпуса локальных документов является актуальной задачей.
CorRIDA насчитывает 1,5 млн. слов, охватывает тексты, адресованные широким категориям пользователей (клиентам), принадлежащие трём социально значимым доменам (здравоохранение, образование, культура), и содержит в том числе разметку по типам текстов. Целью формирования корпуса является, во-первых, описание локальных документов разных типов через выделение и сравнение их языковых черт, во-вторых, оценка официально-деловых текстов с точки зрения их языковой сложности, удобства для восприятия и понимания «простым носителем» русского языка.

Ключевые слова: корпус русских локальных документов, официально-деловые тексты, типы текстов, социально-значимые домены (здравоохранение, образование, культура).

1. Русский официальный документ как объект изучения

В русистике существует традиция описания языка официально-деловых текстов. Прежде всего это работы в рамках стилистики, практической стилистики, «документальной лингвистики», которая в последнее время стала называться «документной лингвистикой», см., например, [1–5].

Компьютерной лингвистикой детальное описание деловых текстов рассматривалось преимущественно в контексте прикладных задач: для успешного решения вопросов технической и деловой коммуникации [6], в том числе — коммуникации человек-машина [7]. Решается задача автоматической кластеризации текстов, в частности, через отнесение к

одному из функциональных стилей, «регистров» или жанров, см., например, [8, 9], обзор см. в [10].

Насколько нам известно, специализированных корпусов русских официально-деловых текстов пока не существует. При этом электронные коллекции документов (репозитории оцифрованных и не оцифрованных текстов на русском языке, относящихся к разным историческим периодам, в том числе — к современности) относительно многочисленны. Упомянем, в частности, собрания нормативных документов на сайте РОМИП (документы Законодательства РФ, Москвы и Санкт-Петербурга, см. [11]), «Полное собрание законов Российской империи» РНБ [12], библиотеку нормативно-правовых актов СССР [13] и др.

Официально-деловые тексты вошли в состав некоторых корпусов русского языка. Так, в «Машинном фонде русского языка» был запланирован текстовой блок «деловая проза» [14]. В Основном корпусе НКРЯ [15] содержатся «нехудожественные тексты», которые можно выбирать, во-первых, по «сфере функционирования», — представлена, в том числе, «официально-деловая» сфера; во-вторых, по «типу текста» (представлены «деловые документы», «законодательные документы», «правовые документы», «судебные документы», «нотариальные документы», внутри каждого типа доступен поиск по жанрам); в-третьих, по «тематике текста» (представлены, в частности, тематические кластеры «администрация и управление» и «право»). Тексты официально-деловой сферы составляют 3,2% от объёма Основного корпуса, см. [16].

Юридические документы, всего 441 текст, в основном — кодексы, входят в Открытый корпус (OpenCorpora), [17]. Некоторое количество деловых текстов (из-за отсутствия жанровой разметки трудно сказать, какое) попало и в состав Russian Business Corpus, одного из корпусов С. Шарова на сайте Университета Лидса [18], и, видимо, в другие русскоязычные веб-корпусы.

Между тем, работ, в которых свойства русских официально-деловых текстов исследовались бы корпусными методами или хотя бы на корпусном материале, крайне мало, см., например, [19]. Это может объясняться, в частности, широчайшим жанровым разнообразием нехудожественных текстов, различиями в применяемых разными авторами типологиях жанров, отсутствием релевантной метаразметки внутри корпусов, а также недостаточной представленностью текстов отдельных жанров в их составе.

Именно в силу «несистематической представленности» текстов официально-делового стиля в существующих ресурсах планируется создание Корпуса официально-деловых текстов русского языка, куда будут включены законы Российской империи, СССР и РФ, императорские указы и постановления советского правительства, см. [20, с. 227].

Формируемый нами Корпус русских локальных документов и актов CorRIDA включает малоисследованную категорию текстов — так называмые **локальные документы** (Internal Documents). Они издаются в конкретной организации или на предприятии администрацией и касаются деятельности только этого предприятия или организации. Для включения в корпус выбраны **документы, адресованные пользователю (клиенту)**: пациенту в поликлинике, абитуриенту в университете и т. д. По-видимому, прежде всего с такими официальными текстами мы (носители русского языка) периодически сталкиваемся: например, читаем и подписываем «Согласия на обработку персональных данных», «Информированные добровольные согласия на медицинское вмешательство», или «Договоры об оказании платных дополнительных услуг».

2. Цели формирования корпуса CorRIDA

Создание корпуса CorRIDA (Corpus of Russian Internal Documents and Acts, Корпус русских локальных документов и актов) производится в рамках исследования, посвящённого функционированию официальных документов в социальных доменах здравоохранения, культуры и образования, подробнее см. [21]. Исследование имеет две магистральные линии, которые можно условно назвать «перцептивной» и

«дескриптивной». В рамках «перцептивного» направления производится анкетирование и интервьюирование носителей русского языка, направленное на выявление доступности официальных документов для восприятия и понимания представителями разных социальных групп.

«Перцептивная» часть исследования начата раньше «дескриптивной». Первоначальным её этапом стало анкетирование ограниченной выборки респондентов и проведение после анкетирования полуструктурированных интервью, см. [21]. Анкеты содержат перечень вопросов к обширным выдержкам из трёх текстов, находящихся в открытом доступе на сайтах учреждений: одного медицинского учреждения (клиники), одного образовательного учреждения (университета) и одного учреждения культуры (музея). Это тексты «Информированного согласия на проведение эндодонтического лечения», «Правил приема в Федеральное государственное бюджетное образовательное учреждение высшего образования» и «Правил поведения» для посетителей музея. Для массового анкетирования мы создали электронные формы анкет.

Перечень вопросов к каждому из трёх перечисленных документов (разных по уровню сложности для восприятия и понимания) направлен, среди прочего, на получение общей оценки текста. Данные массового опроса только предстоит обработать, здесь можно привести некоторые примеры ответов. Так, после выдержки из «Правил приема» следует просьба «*Опишите, пожалуйста, Ваше первое впечатление о тексте. Удобен ли он для чтения? Понятен ли?*».

Большинство респондентов отвечает, что текст в общем понятен (хотя понимание от многих требует усилий), но неудобен для чтения, перегружен, длинен и т. д., ср.: (1) «*слишком громоздко, но понятно*», (2) «*Неудобен, но в целом понятен*», (3) «*Понятен, но перегружен повторами*», (4) «*Все понятно, если читать медленно, но будет ли кто-то это делать? Очень долго, сложно и нудно*», (5) «*неудобен, так как длинные синтаксические конструкции; отчасти понятен, если заставить себя вчитаться. Описания "документов установленного образца" невозможно просто даже дочитать до конца*» и др. Часть респондентов уточняет, что текст будет понятен не всем читателям: (6) «*Нет, не удобен. Понятен более-менее, если уметь выделить в тексте главное. Если русский у человека второй язык (поступающие из союзных республик), то вообще ничего не будет понятно благодаря весёлому согласованию слов в предложениях*», (7) «*Мне да, абитуриенту - вряд ли <понятно> также, как мне*». Некоторые респонденты указывают, что не стали дочитывать фрагмент «Правил приема»: (8) «*многабукафниасилил))) очень длинный, внимание теряется на первых же пунктах*», (9) «*Бросила читать после первой трети, ненужные подробности отвлекают от сути*».

В рамках «дескриптивного» направления исследования планируется многоаспектное лингвистическое описание текстов документов, выполненное корпусными методами. В частности, мы планируем выяснить, насколько сложен для чтения текст официального документа (точнее, определённые типы текстов), опираясь на языковые свойства собранных в корпусе CorRIDA текстов. Для этого мы будем пользоваться и традиционными методами (включающими использование т. наз. «readability formulas», основанными на данных о средней длине предложения или средней длине слова, о частотах слов и пр.), и относительно более новыми методиками оценки языковой сложности (об этом см., в частности, [22, 23]).

Запланировав «дескриптивную» часть исследования, мы столкнулись с нехваткой текстовых ресурсов для её реализации (об официально-деловых текстах в составе русских корпусов см. п. 1 выше). Это привело к мысли о необходимости создания лингвистически размеченного корпуса, содержащего интересующие нас тексты локальных документов.

Общепринятой классификации жанров для нехудожественных текстов не существует, см., например, [8]. Вводятся классификации по разным основаниям, однако разработка жанровой таксономии, основанной на лингвистических свойствах текстов — сложная проблема, которую только предстоит решить, об этом см. [24].

При продумывании состава будущего корпуса мы решили совместить лингвистический и юридический взгляд на документ. Юридическая теория различает, прежде всего, **сделки** (правовые действия, которые порождают конкретные правовые последствия в виде обязательств самих участников сделок) и **правовые акты** (выражают волю уполномоченного лица устанавливать в одностороннем порядке предписания другим лицам). Сделки делятся на **односторонние сделки** (например, завещания) и **договоры** (для которых характерно встречное волеизъявление двух или более лиц). Правовые акты делятся на: **нормативные** (содержат нормы права — общие правила поведения, адресованные неопределенному кругу лиц и рассчитанные на неоднократное применение) и **ненормативные** (содержат конкретные предписания, адресованные конкретным лицам), см. [25].

При формировании корпуса нас интересовали тексты документов, которые можно отнести к категориям **односторонних сделок**, **договоров** и **нормативных правовых актов**. Различение этих категорий значимо и с точки зрения описания композиции и языкового содержания документа. Проиллюстрируем это утверждение одним примером: в текстах односторонних сделок («Согласий на обработку персональных данных») употребительны местоимения 1 л. и глаголы 1 л. ед. ч. («*я … подтверждаю своё согласие*», «*моих персональных данных*», «*предоставляю Оператору право*», «*согласие дано мной*» и т. д.), а в текстах договоров (например, договоров об оказании платных услуг) и разнообразных правил (например, правил поведения пациента) личные и притяжательные местоимения практически не встречаются, контрагенты или лица, чьи права и обязанности оговариваются в документе, поименованы стандартным образом (например, «*Пациент*» и «*Исполнитель*»).

Таким образом, включению в корпус CorRIDA подлежали локальные документы трёх перечисленных категорий (односторонних сделок, договоров и нормативных правовых актов), находящиеся в открытом доступе на сайтах государственных учреждений здравоохранения, образования и культуры (поликлиник, больниц, школ, университетов, музеев, театров и др.).

В конечном счете, нас интересует оценка официально-деловых текстов с точки зрения их языковой сложности, удобства для восприятия и понимания «простым носителем» русского языка (именно в этой точке смыкаются «дескриптивное» и «перцептивное» направления исследования), поэтому в корпус включались только **локальные документы, потенциально адресованные широкому пользователю (клиенту):** пациенту или посетителю в больнице, ученику или родителю в школе, зрителю в театре. К примеру, выбирая между «Кодексом этики и служебного поведения медицинского работника» и «Правилами поведения пациента», мы предпочитали последний тип документа, поскольку он адресован более обширной категории граждан, а не представителям одного профессионального сообщества или коллективу сотрудников конкретного учреждения.

3. Состав корпуса CorRIDA

В корпус вошли тексты односторонних сделок, договоров и нормативных правовых актов. Нас интересовали только **документы, выпущенные государственными учреждениями и адресованные широким категориям граждан**. В результате анализа содержания сайтов учреждений мы выбрали типы документов, которые размещаются на сайтах наиболее регулярно.

В категории так называемых «односторонних сделок» это, прежде всего, «Согласие на обработку персональных данных», а также «Информированное добровольное согласие», встречающееся в доменах здравоохранения. Среди договоров это «Договор об оказании платных услуг». В категории нормативных правовых актов это «Правила поведения (пациента, обучающегося, посетителя)», «Правила оказания платных услуг» и некоторые другие документы, различающиеся по доменам (например, «Правила госпитализации»,

«Правила проведения вступительных испытаний» или «Правила возврата театральных билетов»). Таким образом, мы получили 6 базовых разновидностей локальных документов, отвечающих нашим требованиям. Было решено назвать каждую такую разновидность «типом текста».

Таким образом, в корпус CorRIDA вошли документы, относящиеся к трём социально значимым доменам (образование, здравоохранение, культура), в каждом домене собраны тексты пяти типов. Шестой тип («Информированное согласие») представлен не во всех доменах, поэтому было решено объединить такие тексты с текстами «Согласий на обработку персональных данных».

Поиск и скачивание текстов выполнялось вручную. Такой порядок действий позволил существенно сократить количество времени и усилий, направленных на их предварительную обработку (о стандартных шагах по обработке текстов, собранных с помощью краулеров, см., например, [26]).

Опыт показал, что для нахождения документов удобно пользоваться поисковыми запросами с применением аббревиатур, принятых для обозначения государственных учреждений (ГБУЗ «государственное бюджетное учреждение здравоохранения», ГБОУ «государственное бюджетное образовательное учреждение», ГБУК «государственное бюджетное учреждение культуры», см. также аббревиатуры с префиксами «федеральное», «республиканское», «областное» типа ФГБУЗ, РГБУЗ, ОГБУЗ и др.).

В результате поиска в Интернете с применением кратких названий типов текста и аббревиатур, т.е. при помощи запросов типа «Правила поведения * ГБУЗ» нам удалось собрать текстовую коллекцию размером в 1,5 млн. слов. Состав коллекции описан в таблице 1.

Таблица 1. Состав корпуса в цифрах

домен	тип	кол-во текстов	кол-во слов	среднее значение (медиана)	мин. длина текста в словах	макс. длина текста в словах
Здравоохранение	1	77	110790	1439 (1223)	49	4859
Здравоохранение	2	135	107265	794 (547)	34	3863
Здравоохранение	3	59	101787	1725 (1630)	100	4991
Здравоохранение	4	70	100044	1429 (1362)	439	2766
Здравоохранение	5	153	50337	331.2 (313.5)	27	787
Здравоохранение	6	99	49567	500.7 (362.0)	60	9219
Образование	1	38	105905	2787 (2948)	255	5513
Образование	2	50	104513	2090.3 (1023)	202	13035
Образование	3	51	100432	1969 (1888)	219	4498
Образование	4	73	102318	1402 (1354)	345	3003
Образование	5	258	100973	391.2 (367)	60	1349
Культура	1	106	100861	951.5 (832)	118	3220
Культура	2	62	103312	1666.3 (1487.5)	315	5912
Культура	3	65	100292	1543 (1591)	91	4404
Культура	4	83	100012	1205 (1122)	261	3087
Культура	5	179	45162	252.3 (233)	31	1747
ВСЕГО		**1558**	**1483570**			

Типы текстов обозначены индексами: «1» — Правила поведения (Правила внутреннего распорядка и поведения пациентов, обучающихся и др.); «2» — Порядок (правила) госпитализации, диспансеризации, организации вызова врача, приёма в школу, колледж, университет, Положение о порядке перевода, восстановления, отчисления обучающихся и др.; «3» — Положение об оказании платных услуг (Порядок оказания платных услуг), «4» — Договор на оказание платных услуг (Договор об оказании платных услуг), «5» — Согласие на обработку персональных данных пациента, законного представителя, ребёнка,

родителя и др., «б» — Информированное добровольное согласие (на медицинское вмешательство и др.).

Устойчивые характеристики типов документов, связанные с объёмом в словах и, соответственно, косвенно влияющие на сложность восприятия соответствующих текстов, ещё предстоит описать. Уже сейчас, основываясь на таблице 2, можно заключить, что наиболее протяженными в словах типами текстов в корпусе являются тип «1» в домене «Образование» («Правила внутреннего распорядка обучающихся», «Правила поведения учащихся» и т. п.) и тип «3» в том же домене («Порядок оказания платных образовательных услуг»).

Таблица 2. Ранжирование по убыванию медианных значений длины текстов в словах

домен	тип текста	медиана
Образование	1 (Правила поведения)	2948
Образование	3 (Положение об оказании платных услуг)	1888
Здравоохранение	3 (Положение об оказании платных услуг)	1630
Культура	3 (Положение об оказании платных услуг)	1591
Культура	2 (Порядок (правила) сдачи театральных билетов и др.)	1487
Здравоохранение	4 (Договор на оказание платных услуг)	1362
Образование	4 (Договор на оказание платных услуг)	1354
Здравоохранение	1 (Правила поведения)	1223
Культура	4 (Договор на оказание платных услуг)	1122
Образование	2 (Порядок (правила) отчисления обучающихся и др.)	1023
Культура	1 (Правила поведения)	832
Здравоохранение	2 (Порядок (правила) госпитализации и др.)	547
Образование	5 (Согласие на обработку персональных данных)	367
Здравоохранение	6 (Информированное добровольное согласие)	362
Здравоохранение	5 (Согласие на обработку персональных данных)	313
Культура	5 (Согласие на обработку персональных данных)	233

4. Структура корпуса CorRIDA

Корпус будет состоять из трёх подкорпусов по доменам (домены «здравоохранение», «образование», «культура») и шести подкорпусов по типам текстов. Так как коллекция, которая станет основой корпуса, достаточно однородна по составу, представляется, что минимального набора метаданных для описания текстов в её составе достаточно. Каждый документ внутри коллекции сопровождается следующим набором сведений:

- название домена;
- название учреждения, с сайта которого получен документ;
- название документа;
- стандартное название типа документа;
- условный номер документа в субколлекции;
- источник электронной версии (адрес сайта в Интернете, с которого документ был скачан);
- дата скачивания документа.

Пользователь корпуса будет иметь доступ к информации о домене, типе документа, названии документа.

Документы скачивались вручную, поэтому предварительная обработка текстов для включения в корпус была несложной и заключалась, прежде всего, в удалении подряд идущих пробелов и табуляций, удалении пустых строк, удалении строк, содержащих только пробелы или табуляции, удалении пробелов в начале строк, замене парных кавычек на прямые и т. п. Дедупликация на уровне целых документов не требуется, так как тексты, вошедшие в корпус, оценивались экспертами (дубликаты в корпус не включались). Проверка на наличие нечетких дубликатов на уровне компонентов документа не

планируется, так как нам интересны, в том числе, повторяющиеся элементы, способные показывать строгость соблюдения в конкретном тексте определенного шаблона.

Вопросом, потребовавшим решения, стал вопрос об анонимизации данных. В корпусной практике анонимизации подвергаются различные персональные данные, в первую очередь — фамилии, адреса, телефонные номера и т. д.

В нашем корпусе собраны документы, размещённые в Интернете в открытом доступе. Тем не менее, поскольку в дальнейшем мы, во-первых, планируем сделать корпус общедоступным, предоставив желающим возможность его скачивания, во-вторых, намереваемся оценивать документы с точки зрения их языковых качеств, мы приняли решение выполнить анонимизацию некоторых личных данных.

Существует два основных способа анонимизации:

1. Удаление данных [27, p. 62].

2. Замена данных, например, замена имён псевдонимами, буквенными или цифровыми кодами. О случаях, когда замены необходимы или нежелательны, см. [28, p. 14–19]. В частности, правила замен для фамилий таковы: псевдонимы должны быть фонетически и просодически похожи на оригинальные имена, содержать одинаковое число слогов, должны начинаться с тех же букв, что и оригинальные имена.

Можно с уверенностью утверждать, что в документе имена собственные не обыгрываются ни фонетически, ни ритмически, ни в ходе языковой игры, поэтому отображать их первоначальный облик в псевдонимах не обязательно. Соответственно, выбран способ анонимизации, при котором названия государственных учреждений, почтовые и электронные адреса, телефоны и пр. частично преобразуются, при этом количество слов в тексте сохраняется неизменным. В результате получаем наименования и фамилии типа: «*ГБУЗ NК "Nская ЦРБ"*», «*КГАУЗ "Nская стоматологическая поликлиника"*», «*портал государственных услуг N-ского карая* (sic) *(https://uslugi00.ru/)*», «*записаться на прием можно лично у секретаря главного врача либо по телефону: 000-00-00*», «*Nкину Константину Владимировичу*», «*Nкиной Ирине Николаевне*» и пр. Таким образом, мы легко можем определить место замены и общее значение слова, аббревиатуры, алфавитно-цифрового комплекса и т. п., подвергшегося преобразованию.

5. Заключение

В повседневной жизни носители русского языка всё чаще сталкиваются с необходимостью читать и подписывать различные официальные документы. Обычно это так называемые локальные документы (Internal Documents). Мы встречаемся с ними, обращаясь за медицинской помощью, оформляя Шенгенскую визу или возвращая в кассу театральные билеты. Между тем, язык таких документов исследован недостаточно и практически не рассматривался с применением корпусных методов.

Существующие корпусы русского языка пока не предоставляют возможностей для систематического анализа языка документа. Это связано, в частности, с проблемами жанровой классификации и разметки нехудожественных текстов. Поэтому мы решили сформировать лингвистически размеченный Корпус русских локальных документов и актов CorRIDA. Этот корпус позволит выполнить описание локальных документов разных жанров через выделение и сравнение их языковых черт.

Исследование выполняется в рамках НИР по анализу соблюдения норм современного русского литературного языка при его использовании в качестве государственного в деятельности организаций культуры, здравоохранения и образования, включённой в План мероприятий НИИ Проблем государственного языка СПбГУ во исполнение Комплекса мер, направленных на совершенствование государственной политики в области развития, защиты и поддержки русского языка на 2016-2020 гг. Совета по русскому языку при Правительстве РФ.

Литература

[1] Рахманин Л.В. Стилистика деловой речи и редактирование служебных документов. Учебное пособие. М.: Высшая школа, 1988.

[2] Шварцкопф Б.С. Официально-деловой язык // Культура русской речи и эффективность общения. М., 1996. С. 270 – 281.

[3] Дюженко Г.А. Документальная лингвистика. М., 1975.

[4] Янковая В.Ф. Документная лингвистика. М.: Издательский центр «Академия», 2011.

[5] Муравьёва М.Н. Документная лингвистика. М.: Издательство «ТЕРМИКА», 2016.

[6] Герд А.С. Предмет и основные направления прикладной лингвистики // Прикладное языкознание. СПб., 1996. URL: http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html (дата обращения: 27.01.2018).

[7] Ершов А.П. К методологии построения диалоговых систем. Феномен деловой прозы. Новосибирск, 1979.

[8] Пиперски А.Ч. Жанровая классификация в Генеральном интернет-корпусе русского языка // Современные проблемы науки и образования. 2013. № 4. URL: https://www.science-education.ru/ru/article/view?id=9762 (дата обращения: 27.01.2018).

[9] Дубовик А.Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам // Компьютерная лингвистика и вычислительные онтологии. Выпуск 1 (Труды XX Международной объединенной научной конференции «Интернет и современное общество, IMS-2017, Санкт-Петербург, 21 - 23 июня 2017 г. Сборник научных статей). СПб: Университет ИТМО, 2017. С. 29 – 45. URL: http://openbooks.ifmo.ru/ru/file/6502/6502.pdf (дата обращения: 27.01.2018).

[10] Conrad S. Register variation // Biber D., Reppen R. (eds.) The Cambridge handbook of English corpus linguistics. Cambridge University Press, 2015. P. 309 – 329.

[11] РОМИП (Российский семинар по оценке методов информационного поиска). URL: http://romip.ru/index.html (дата обращения: 27.01.2018).

[12] Полное собрание законов Российской империи. URL: http://www.nlr.ru/e-res/law_r/about.html (дата обращения: 27.01.2018).

[13] Библиотека нормативно-правовых актов СССР. URL: http://www.libussr.ru/ (дата обращения: 27.01.2018).

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

[15] Национальный корпус русского языка. URL: http://www.ruscorpora.ru/ (дата обращения: 27.01.2018).

[16] Статистика корпуса. URL: http://www.ruscorpora.ru/corpora-stat.html (дата обращения: 27.01.2018).

[17] «Открытый корпус» (OpenCorpora). URL: http://opencorpora.org/ (дата обращения: 27.01.2018).

[18] Russian Business Corpus. URL: http://corpus.leeds.ac.uk/ruscorpora.html (дата обращения: 27.01.2018).

[19] Буторина Е.П. Категория официальности в современном русском языке. Автореф. дисс. … докт. филол. наук. М., 2016.

[20] Крылов С.А., Фролова О.Е. О корпусе официально-деловых текстов русского языка // Труды международной конференции «Корпусная лингвистика-2017». СПб, 2017. С. 226 – 230.

[21] Гулида В.Б. Социолингвистическая проблематика официальных документов // Социо- и психолингвистические исследования. Вып. 4. 2016. С. 112–125. URL: https://elibrary.ru/item.asp?id=28381522 (дата обращения: 27.01.2018).

[22]Hancke J., Vajjala S., Meurers D. Readability classification for German using lexical, syntactic, and morphological features // Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). Mumbai, India, 2012. P. 1063–1080.
[23]Crossley S., Dufty D., McCarthy Ph., McNamara D. Toward a new readability: A mixed model approach // Proceedings of the 29th annual conference of the Cognitive Science Society. Nashville, Tennessee, USA, 2007. P. 197–202.
[24]Кибрик А.А. Анализ дискурса в когнитивной перспективе. Дис. … д-ра филол. наук. М.: Ин-т языкознания РАН. 2003. URL: http://iling-ran.ru/kibrik/DA_cognitive_perspective@Diss_2003.pdf (дата обращения: 27.01.2018).
[25]Бошно С.В. Развитие признаков нормативного правового акта в современной правотворческой практике // Журнал российского права. 2004. №2. С. 95 – 106. URL: https://elibrary.ru/item.asp?id=26350658 (дата обращения: 27.01.2018).
[26]Barbaresi A. Ad hoc and general-purpose corpus construction from web sources. Linguistics. ENS Lyon, 2015. URL: https://tel.archives-ouvertes.fr/tel-01167309/document (дата обращения: 27.01.2018).
[27]McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012.
[28]Hasund K. Protecting the innocent: the issue of informants' anonymity in the COLT corpus // Explorations in Corpus Linguistics. Amsterdam: Rodopi, 1998. P. 13 – 28.

Corpus of Russian Internal Documents and Acts CorRIDA: Goals, Composition and Structure

S.A. Belov [1], O.V. Blinova [1], V.B. Gulida [1], V.Yu. Zubov [1],
E.Yu. Larionova [2], P.S. Tolstikova [3]

[1] St. Petersburg State University, [2] European University at Saint-Petersburg,
[3] Institute for Linguistic Studies, Russian Academy of Sciences

The existing Russian corpora do not yet provide opportunities for a systematic analysis of the language of official documents. There are few such texts in existing corpora. Moreover, there are the problems of genre classification and markup of non-fiction (incl. official, legal) texts.
The paper describes the initial creation stage of the corpus of Russian Internal Documents and Acts «CorRIDA». In everyday life, Russian speakers are increasingly faced with the need to read and sign various official documents. Usually these are so-called «internal documents», for example, Contracts or Informed Consents. However, the language of such documents has not been examined with the use of corpus methodology.
The corpus contains 1.5 million words, includes documents belonging to three socially significant domains (health, education, culture) and will allow the description of internal documents of various types.

Keywords: Legal Corpora, Official Texts, Corpus of Russian Internal Documents, Socially Important Domains