Шерстинова Т. Ю., Гамзатова А. Ф.
Sherstinova T. Yu., Gamzatova A. F.

О некоторых возможностях компьютерной лингвистики для междисциплинарных исследований права, языка и литературы

On some possibilities of computer linguistics for interdisciplinary studies of law, language and literature

Abstract: The article presents three linguistic resources which can be used for interdisciplinary research of law, language and literature. The online service of Google Books Ngram Viewer makes it possible to trace the dynamics of legal terms usage for more than 200 years. The similar online opportunities for analyzing Russian data are provided by the National Corpus of the Russian Language, which is the largest collection of Russian texts (both fiction and non-fiction). Using different filters (temporary, genre, thematic, etc.), we can study the frequency of occurrence of certain legal terms, as well as other concepts related to jurisprudence, in texts by different authors and get an extended context of their usage. Finally, the biographical database of Russian prose writers of the first third of the 20th century, which is being created as a module of the Corpus of Russian Short Stories of 1900–1930, provides information on who of the writers were lawyers by education and worked in their specialty.

Keywords:
Ключевые слова:

Компьютерная лингвистика — относительно молодая и быстро развивающаяся область знания, одной из основных целей которой является автоматизация разного рода задач, связанных с анализом текста [1]. Уже само название указывает на междисциплинарность подходов, которые предполагают использование строгих формальных методов (в первую очередь, математических и статистических) в применении к тексту — объекту, традиционно изучаемому филологией, то есть сугубо гуманитарной дисциплиной.

С повсеместным использованием персональных компьютеров, которые уже есть фактически в каждом доме, и развитием интернета, количество циркулирующих в информационном пространстве текстов постоянно возрастает. Помимо традиционно существующих библиотек, которые постепенно переводятся в цифровую форму, и электронных изданий (литературных, публицистических, новостных и т. д.), регулярно появляются новые коллекции текстов — специально разрабатываемые лингвистические корпуса, а также «стихийно» создающиеся интернет-порталы, содержащие массивы постоянно разрастающихся web-страниц, блогов, интернет-форумов, социальных сетей, наполненных текстовой

информацией по самым разным отраслям знания и человеческой деятельности.

Современные информационные технологии позволяют осуществлять поиск нужной информации среди этого поистине бескрайнего множества текстов. В некоторых случаях для этого удобно использовать специально подготовленные ресурсы — лингвистические корпуса, базы данных и др. Объектом исследования во всех случаях будет или сам текст и его особенности, или особенности языка данного типа текстов.

В этой работе, имеющей скорей популярный, нежели научный характер, мы хотим показать, как можно использовать лингвистические ресурсы для проведения отдельных исследований, относящихся к такой, казалось бы далекой от филологии, дисциплине, как право.

Поскольку семинар, по материалам которого подготовлены эти статьи, посвящен междисциплинарным исследованиям права, языка и художественной литературы, коснемся и литературного аспекта этого взаимодействия. Художественные тексты по понятным причинам давно находятся в фокусе интересов компьютерной лингвистики [2], а также предоставляют богатый материал и для исследований в области права (в первую очередь в историческом аспекте), языка права и юридической терминологии.

Итак, какие возможности компьютерной лингвистики можно было бы порекомендовать исследователям, занимающимся изучением языка права (как впрочем, и многих других дисциплин) и соответствующей терминологии?

1. Google Books Ngram Viewer

Наглядный пример использования лингвистического интернет-ресурса для диахронического исследования ключевых юридических понятий и частоты их употребления демонстрирует Google Books Ngram Viewer. Это поисковый онлайн-сервис компании Google, который позволяет строить графики частотности языковых единиц на основе внушительного корпуса печатных источников, опубликованных с начала XIX века вплоть до недавнего времени (точнее, до 2008), представленных в крупнейшем мировом собрании цифровых книг Google Books [3].

С помощью этого онлайн сервиса можно определять частотность употребления слов и словосочетаний в различные временные периоды (1800–2008 гг.). Тем самым становится возможным проследить динамику юридических терминов за более чем двухсотлетний период. Крайне важным является факт, что получаемые в результате запроса данные не представляются в виде набора «сухих» цифр, а подкрепляются онлайн

примерами из реально существовавших нормативно-правовых актов, а также текстов научного и художественно-публицистического жанров.

Рассмотрим для примера динамику использования таких ключевых понятий юриспруденции, как «право», «закон», «договор». После выполнения запроса на корпусе русскоязычных источников, учитывающего как современную, так и дореволюционную орфографию, получаем следующие данные (см. рис. 1). Из рисунка видно, что за прошедшие 200 лет наиболее популярным юридическим термином из рассмотренных было «право», причем наблюдаются очевидные пики его употребления с максимумом в районе 1860 г., что, очевидно, связано с отменой крепостного права в 1961 г.

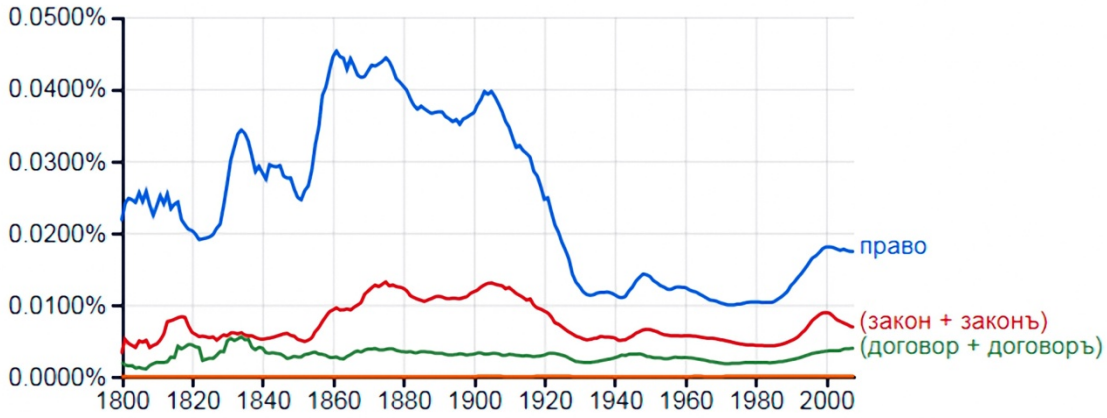

Рис. 1. Сравнительная динамика частоты использования юридических терминов «право», «закон», «договор»

Возможности сервиса позволяют определить частотность не только отдельно встречающихся слов, но и их сочетаемость с другими знаменательными и служебными частями речи. Так, сочетание «право собственности» выступает наиболее частотным в 1845–1860 годах и находит отражение в журналах «Отечественные записки» (в статье о законах литературной собственности в Западной Европе — выпуск за январь 1853 года), «Современник» (в опубликованном там сочинении Александра Лакиера «О вотчинах и поместьях» — выпуск за 1848 год), «Полном собрании законов Российской империи» (том XXVI, 1854 год) и ряде других документов научно-популярного, художественного, и официально-делового стиля.

Кроме того, при помощи этой программы становится возможным установить год зарождения термина или приобретения им нового значения. Например, как выяснилось, термин «правопорядок» впервые появился в конце XIX века в русскоязычном переводе поэмы Джона Мильтона «Потерянный Рай» за 1871 год и стремительно набирал

популярность к концу XX века. Обладание сведениями такого характера позволяет юристам иметь наиболее полную картину о правовой системе предшествующих эпох и сравнивать устойчивость и изменчивость понятий в диахронии.

По частоте употребления тех или иных имен в книгах и иных печатных изданиях можно оценить степень популярности того или иного деятеля (политика, ученого, писателя, художника, композитора и др.). Проверим это на примере трех значительных фигур в области юриспруденции — голландского юриста XVII века Гуго Гроция и выдающихся русских юристов XIX века —Фёдора Плевако и Анатолия Кони.

Графики, построенные программой, наглядно демонстрируют, что имя основоположника международного права Гуго Гроция (Hugo Grotius, Hugo de Groot) в мировой литературе последнего столетия встречается значительно чаще, чем упоминание самых известных юристов Российской империи. Впрочем, этот факт вполне предсказуем. Однако с помощью Google Books Ngram Viewer, мы можем сравнить интерес к этой выдающейся личности в разных странах Европы и в разные временные периоды. Оказывается, что интерес к наследию и личности Гуго Гроция сильнее всего выражен в литературе на немецком и голландском языках, причем пик этой популярности приходится на 1869–1870 гг., когда на немецком языке были изданы труды этого великого юриста.

Сравнивая частоту употребления имен Анатолия Кони и Фёдора Плевако в русскоязычных текстах, можно заметить, что несмотря на то, что А. Кони при жизни был более известен, чем его современник Ф. Плевако, о чём свидетельствуют не только исторические факты, но и показатели графика, в настоящее время уровень цитируемости его трудов и упоминание его фамилии в научных изданиях значительно сократились, в то время как интерес к работам Плевако, напротив, к концу 20-го столетия возрос. Использование более сложных поисковых запросов позволяет получить и другие интересные данные.

2. Лингвистические корпуса и корпусная лингвистика

Подобным образом — но только русском материале — можно исследовать употребление юридических понятий и частоту встречаемости имен отдельных персоналий с помощью Национального корпуса русского языка (НКРЯ), который представляет собой крупнейшее собрание русских текстов разных жанров (художественных, научных, публицистических и др.) [4]. Поиск информации осуществляется онлайн по ключевым словам, задаваемым пользователем. Используя разные фильтры (временные, жанровые, тематические и пр.) можно проследить частоту появления тех

или иных юридических терминов и других релевантных понятий у разных авторов и получить расширенный контекст их употребления.

Результаты поисковых запросов представляют традиционную юридическую терминологию во всём многообразии контекстного употребления. Это облегчает её понимание для неспециалистов (в частности, лингвистов), не имеющих специального юридического образования, и позволяет отследить синтагматические связи слов в зависимости от контекста аналогичных правовых ситуаций.

Подобно рассмотренному выше сервису от Google, НКРЯ предоставляет статистику о частоте употребления искомого слова-термина и контексте его употребления. Преимуществом поиска по корпусу является возможность задания сложных запросов, учитывающих словоизменительную парадигму (поиск по лемме), синтаксические связи, словообразование, семантику и много других лингвистических параметров, а также метатекстовые данные об источнике.

В качестве примера покажем, какую информацию можно получить в результате поиска в Основном корпусе для юридического термина «санкция». Уточним, что поиск производится по всем словоформам и осуществляется с помощью «старой» версии поиска, активно используемой до сентября 2019 г. Установив фильтр по нехудожественным текстам, мы получаем 1293 документа, в которых встречается этот термин общей частотой 2347 вхождения. Частота использования леммы «санкция» для периода 1900–2014 гг. представлена на графике на рис. 2.

Можно видеть, что максимум частоты употребления данного термина приходится на 2014 год, что в полной мере отвечает сложившейся в то время мировой политической обстановке. Статистика НКРЯ говорит также о том, что чаще всего этот термин встречается в статьях (54%), монографиях (8%) и мемуарах (8%). Можно получить данные и о тематическом распределении текстов, содержащих данное слово (так, лишь в 4,26% случаев этот термин присутствует в текстах исключительно правовой тематики, в 25,55% он используется в текстах, посвященных политике и общественной жизни и т. д.).

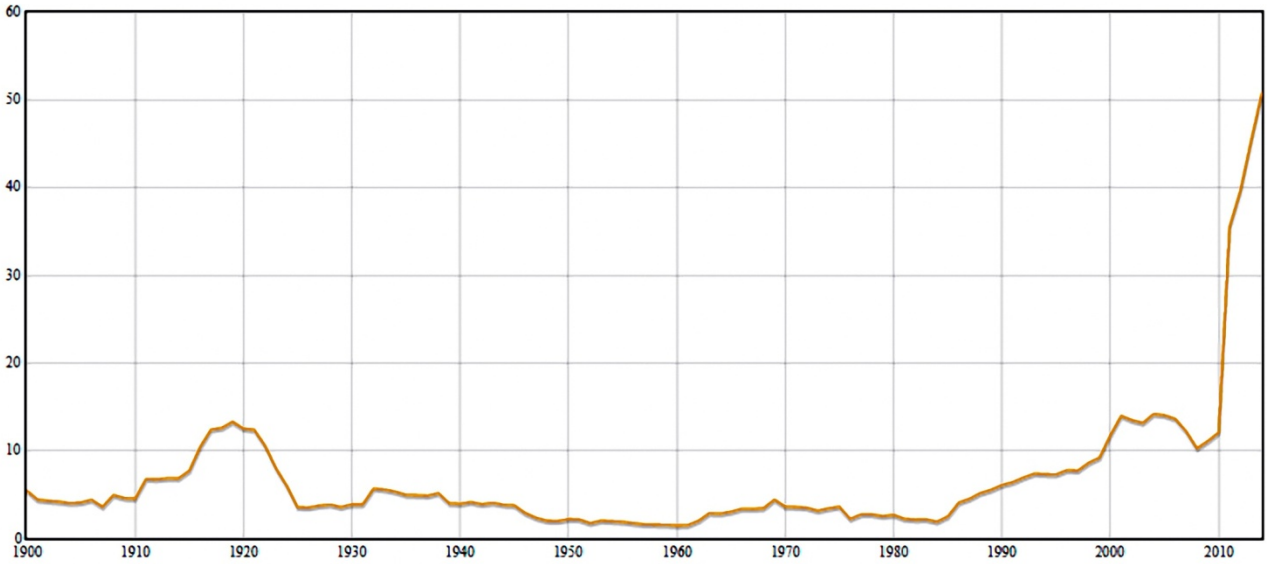

Рис. 2. Частота употребления слова «санкция» в диахронии для нехудожественных текстов с 1900 по 2014 г.

Следует отдельно отметить важность крупных лингвистических корпусов (НКРЯ и других) как реферативных ресурсов для проведения лингвистической экспертизы текстов. Экспертная работа с текстовыми данными (лингвистическая экспертиза) — относительно новое направление в юридической практике, связанное с анализом и интерпретацией текстов, которые привлекаются к ведению дел, связанных с защитой чести и достоинства, разжиганием межнациональной и межрелигиозной розни, публичными оскорблениями, защитой товарных знаков и т. д. Как справедливо отмечает известный специалист в области лингвистической экспертизы А. Н. Баранов: «В отличие от обычного лингвистического исследования, требование обоснованности для экспертизы оказывается не только одним из принципов научности, но и основным условием ее проведения, поскольку ее результаты могут быть положены в основу судебного решения. ... Иными словами проблема использования языковых данных для лингвистической экспертизы текста оказывается столь же актуальной, как и для задач теоеретической лингвистики» [5, с. 475]. С этой точки зрения репрезентативные лингвистические корпуса предоставляют экспертам важную статистическую информацию, необходимую для определения значения свободного словосочетания, для оценки типичности употребления слова или словосочетания в том или ином значении, для раскрытия эллиптичного варианта устойчивого выражения, для экспертизы семантики текста, для регистрации и охраны товарных знаков и др. Подробнее о методах лингвистической экспертизы см. [5].

3. Биографическая база данных русских прозаиков

Данный цифровой ресурс создается как независимый модуль Корпуса русского рассказа первой трети XX века — уникального собрания художественных текстов, предназначенного для многомерного статистического описания языка судьбоносной для России эпохи и моделирования ее литературно-художественных систем (дореволюционной, революционной и постреволюционной) [6]. При создании корпуса русского рассказа первой трети XX века ставится задача включения в текстовую базу максимального числа авторов (а не только «школьных классиков» Толстого, Бунина, Чехова), творивших в исследуемую эпоху, тем самым должна повыситься объективность проводимых на нём лингвостилистических исследований.

Для изучения особенностей авторского стиля и языка отдельных писателей кажется целесообразным принимать во внимание определенные аспекты их биографии и социального происхождения. По-видимому, не последнюю роль в формировании стилистических навыков и мировоззренческих предпочтений писателя играют, в частности, воспитавшая его социальная среда, образование и профессия. Учет этих факторов при квантитативном анализе художественных текстов позволит определить, действительно ли существует зависимость между социальными характеристиками автора (и некоторыми важными особенностями его биографии) и индивидуальностью его языка и стиля, и какова ее мера [7].

Помимо того, представленная в базе данных информация может дать обобщенную картину русского литературного пространства той поры. Например, какова в начале XX века была доля писателей дворянского происхождения, крестьянского происхождения, «разночинцев» и т. д. В контексте тематики данной статьи нам было интересно получить информацию о том, кто из русских писателей был дипломированным юристом и имел практику работы по специальности. На основании 267 биографий писателей, представленных в базе к настоящему времени, оказалось, что писатели-юристы в русской литературе действительно присутствуют, и среди них довольно много известных имен. Так, мало кто знает, что опыт юридической работы (пусть и не всегда продолжительный) имели Л. Н. Андреев, Б. А. Лазаревский, И. С. Шмелев, Ф. А. Червинский, М. П. Чехов, Н. П. Ашешов, В. А. Мазуркевич, С. Д. Кржижановский и некоторые другие. Достаточно представителен и круг авторов, которые учились на юридических курсах, но по разным причинам их не закончили. Это Л. Н. Толстой, Д. Н. Мамин-Сибиряк, Н. Г. Гарин-Михайловский, Е. Н. Чириков и др.

Полученные сведения могут использоваться для решения разных задач как литературоведами, так лингвистами и юристами. Например,

интересной междисциплинарной задачей является изучения языковых особенностей писателей, совмещавших творческую деятельность с юридической практикой. Так, интересно было бы рассмотреть, насколько казенный язык нормативно-правовых актов влияет на художественный текст и как это отражается на особенностях художественного произведения. Надо сказать, что для современной устной речи юристов такие исследования проводились, и отдельные «профессиональные черты» в ней определенно просматриваются [8].

Статья подготовлена при поддержке гранта РФФИ № 17-29-09173 «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

Литература

[1] Прикладная и компьютерная лингвистика / Под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. — М.: ЛЕНАНД, 2016.
[2] Мартыненко Г. Я. Методы математической лингвистики в стилистических исследованиях. — СПб: Нестор-История, 2019.
[3] Google Books Ngram Viewer, https://books.google.com/ngrams (дата обращения 30.09.2019)
[4] Национальный корпус русского языка, http://www.ruscorpora.ru (дата обращения 30.09.2019)
[5] Баранов А. Н. Технология экспертной деятельности: корпусы текстов в лингвистической экспертизе текста / Баранов А. Н. Лингвистическая экспертиза текста. М.: ФЛИНТА: Наука, 2013.
[6] Мартыненко Г. Я., Шерстинова Т. Ю., Попова Т. И., Мельник А. Г., Замирайлова Е. В. О принципах создания корпуса русского рассказа первой трети XX века // Труды XV Международной конференции по компьютерной и когнитивной лингвистике «TEL 2018». – Казань, 2018. – С. 180–197.
[7] Шерстинова Т. Ю. Биографическая база данных русских писателей (к созданию корпуса русского рассказа XX века) // Труды международной конференции «Корпусная лингвистика-2019». — СПб.: Изд-во С.-Петерб. ун-та, 2019, с. 439–447.
[8] Куканова В. В. Лингвистический анализ репродуцированных текстов: на материале звукового корпуса русской речи юристов: дис. ... канд. филол. наук; СПбГУ, Санкт-Петербург, 2009.