

## РАЗРАБОТКА ЛИНГВИСТИЧЕСКОЙ БАЗЫ ДАННЫХ ЮРИДИЧЕСКОГО ДОМЕНА В ФОРМАТЕ NOOJ

**К. С. Кисель**

*Белорусский государственный университет, пр. Независимости, 4,  
220030, г. Минск, Беларусь, ksenia.kisel126@gmail.com  
Научный руководитель – В. В. Воронович, старший преподаватель*

В статье рассматривается понятие «лингвистическая база данных» и описывается специфика разработки такой базы в юридическом домене. Также представлен и проанализирован корпус кодексов Республики Беларусь, собранный в программе NooJ.

**Ключевые слова:** лингвистическая база данных; NooJ; база данных; корпус текстов; машинный словарь; юридическая терминология.

В современном мире информационных технологий базы данных играют ключевую роль в обработке и хранении разнообразной информации. База данных (БД) – это «упорядоченная совокупность данных, предназначенных для хранения, накопления и обработки с помощью ЭВМ» [1, с. 96]. Современные базы данных позволяют хранить не только данные, но и методы, которые управляют взаимодействием с пользователем или другими программно-аппаратными комплексами.

Лингвистическая база данных (ЛБД) – это собрание информации, связанной с языком, которое может включать словари, тексты, аудио- и видеозаписи, а также другие данные, относящиеся к языку и его использованию. В юридическом домене разработка такой базы данных может быть особенно полезной для лингвистических экспертиз, например, при анализе юридической документации, исследовании отрицательно-оценочной лексики или для поддержки работы судебных переводчиков [2].

Лингвистическая база данных в юридическом домене, особенно в формате NooJ, представляет собой ценный инструмент для анализа и обработки юридического языка. NooJ – это гибкая компьютерная платформа, которая позволяет лингвистам разрабатывать и использовать сложные лингвистические ресурсы, такие как электронные словари, грамматики и корпуса текстов.

При создании ЛБД и определении задач её функционирования следует в первую очередь учитывать такие специфические черты языка права, как точность, однозначность, интегрированность, унифицированность, обобщенность, а также формализованность правовой терминологии [3].

Для успешной разработки лингвистической базы данных юридического домена нужно принять во внимание использование соответствующих эффективных подходов и технологий для обработки юридической информации. Рассмотрим некоторые из них.

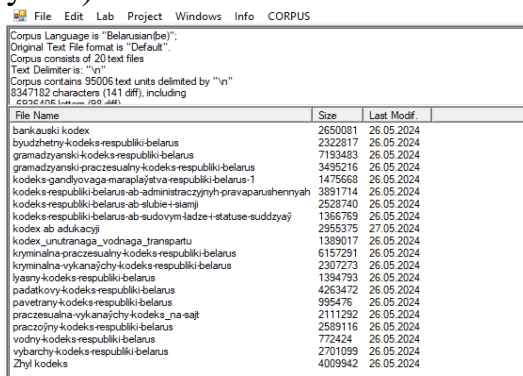
1. **Прагматически-ориентированный подход.** В рамках данного подхода акцент делается на практические аспекты использования ЛБД в юридической деятельности. Прагматически-ориентированный подход сосредоточен на выполнении конкретных задач, таких как поиск информации, анализ текстов и обработка данных.

2. **Использование правовых систем.** Например, такие системы, как «Гарант» и «КонсультантПлюс», предоставляют доступ к нормативно-правовым документам, практике суда и другим юридическим документам [4].

3. **Информационные технологии в Интернете.** Юристы работают с большими объемами данных, поэтому часто используют программы для работы с информацией, сегодня трудно представить юридические дела без информационных технологий. Поиск и анализ данных в Интернете позволяют юристам получать актуальную информацию.

Анализ корпусов юридических текстов представляет собой важный этап в изучении и понимании правовой системы. Корпусы текстов содержат большое количество юридических документов, которые могут быть проанализированы для выявления закономерностей, тенденций и особенностей правового языка. О.Г. Горина определяет в своем исследовании компьютерный языковой корпус как «совокупность языковых данных, представленных в электронном виде с возможностью сплошного анализа материала и автоматического вывода результатов» [5, с. 1]. Поиск в корпусе позволяет создать конкорданс, который включает в себя список каждого употребления слова в контексте, а также ссылки на источники [6, с. 12].

Рассмотрим процесс разработки лингвистической базы данных юридического домена с использованием инструмента NooJ [15]. Данный корпус включает в себя кодексы Республики Беларусь, переведённые на белорусский язык (рисунок).



The screenshot shows the NooJ Corpus window. At the top, it says 'Corpus Language is "Belarusian(bel)"; Original Text File format is "Default"; Corpus consists of 20 text files; Text Delimiter is: "\n"; Corpus contains 95006 text units delimited by "\n"; 8347182 characters (141 diff), including 6832 diff tokens (260 diff)'. Below this is a table with three columns: File Name, Size, and Last Modif.

| File Name                                                      | Size    | Last Modif. |
|----------------------------------------------------------------|---------|-------------|
| bankauski kodeks                                               | 2650081 | 26.05.2024  |
| byudzhetny-kodeks-respubliki-belarus                           | 2322817 | 26.05.2024  |
| gramadzyanski-kodeks-respubliki-belarus                        | 7193483 | 26.05.2024  |
| gramadzyanski-praczesualny-kodeks-respubliki-belarus           | 3495216 | 26.05.2024  |
| kodex-gandyovaga-marapajstva-respubliki-belarus-1              | 1475668 | 26.05.2024  |
| kodex-respubliki-belarus-ab-administracyjnyh-pravaparushennyah | 3891714 | 26.05.2024  |
| kodex-respubliki-belarus-ab-sluzbie-sienij                     | 2528740 | 26.05.2024  |
| kodex-respubliki-belarus-ab-sudovym-ladze-i-statuse-sudzyaj    | 1366769 | 26.05.2024  |
| kodex-ab-adukacyj                                              | 2955375 | 27.05.2024  |
| kodex-unutrana_vodnaga_transportu                              | 1389017 | 26.05.2024  |
| kryminalna-praczesualny-kodeks-respubliki-belarus              | 6157231 | 26.05.2024  |
| kryminalna-vykanaicny-kodeks-respubliki-belarus                | 2307273 | 26.05.2024  |
| lyamy-kodeks-respubliki-belarus                                | 1394793 | 26.05.2024  |
| padatkovy-kodeks-respubliki-belarus                            | 4263472 | 26.05.2024  |
| pavetranny-kodeks-respubliki-belarus                           | 995476  | 26.05.2024  |
| praczesualna-vykanaicny-kodeks-na-sajt                         | 2111292 | 26.05.2024  |
| praczojny-kodeks-respubliki-belarus                            | 2589116 | 26.05.2024  |
| vodny-kodeks-respubliki-belarus                                | 772424  | 26.05.2024  |
| vybarchy-kodeks-respubliki-belarus                             | 2701099 | 26.05.2024  |
| Zhnyj kodeks                                                   | 4009942 | 26.05.2024  |

Корпус юридических кодексов в формате NooJ

Проведем лингвистический анализ данного корпуса, рассмотрев частотность юридических терминов и выражений.

Проведя анализ данного корпуса, можно сделать вывод о частом употреблении стандартных фраз и терминов, которые описывают правовые нормы, обязанности, права и процедуры. Например, самое часто встречающееся словосочетание *Рэспублікі Беларусь* указывает на государственную принадлежность, а *гэтага Кодэкса* относится к конкретному законодательному акту. Фразы *ў адпаведнасці, мае права, за сабой* и другие используются для согласования предложений и создания логической структуры текста. Они помогают выразить связь между разными частями законодательных актов. Кроме того, столь частые упоминания *etalonline by, regnum hk, document regnum* и *by document* связаны с ссылками на другие юридические документы и источники, которые уточняют и раскрывают определенные аспекты. В целом, частое использование этих словосочетаний связано с особенностями юридического стиля и необходимостью точности и ясности в правовых текстах.

Таким образом, объем и разнообразие данных в лингвистических базах данных юридического домена играют ключевую роль в обеспечении глубины и широты анализа юридической информации, а также в поддержке многоязычных и международных правовых исследований, а созданная лингвистическая база данных юридического домена позволит исследователям и специалистам в области юриспруденции эффективнее анализировать юридические тексты, выявлять закономерности и обогащать лингвистические ресурсы. Изучение корпуса поможет выявить общие структурные черты юридических документов, такие как типичные заголовки, разделы, а также структуру предложений. Собранный корпус может быть полезен для автоматической обработки и анализа юридических текстов, разработки вопросно-ответных систем и других приложений. Кроме того, для составления юридических словарей и определения ключевых понятий будет полезен анализ частотности слов с целью определения наиболее часто встречающихся терминов и фраз.

### Библиографические ссылки

1. Прохорова А.М. База данных // Большой энциклопедический словарь / М. Сов. энциклопедия, 1993. С. 96.

2. Кочергина К. С. Лингвистическая база данных отрицательно- оценочной лексики: концепция, структура, наполнение [Электронный ресурс] // Вестн. Том. гос. ун-та. 2019. №446. URL: <https://cyberleninka.ru/article/n/lingvisticheskaya-baza-dannyh-otritsatelno-otsenочноy-leksiki-kontseptsiya-struktura-napolnenie> (дата обращения: 29.05.2024).

3. *Виландеберк, А. А.* Корпус параллельных правовых документов как составная часть АРМ юриста-переводчика [Электронный ресурс] // Труды международной конференции „Корпусная лингвистика – 2004“. – Санкт-Петербург: Изд-во Санкт-Петербургского ун-та. – С. 72-85. URL: [https://events.spbu.ru/eventsContent/files/corpling/corpora2004/Vilandeberk\\_art.pdf](https://events.spbu.ru/eventsContent/files/corpling/corpora2004/Vilandeberk_art.pdf) (дата обращения: 29.05.2024).

4. *Информационные технологии в юридической деятельности* [Электронный ресурс]. URL: <https://mosdigitals.ru/media/informatsionnye-tekhnologii-v-yuridicheskoy-deyatelnosti> (дата обращения: 29.05.2024).

5. *Горина О.Г.* Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов в профессионально-ориентированном общении на английском языке: автореф. дис. ... канд. пед. наук. - М., 2014. - 27 с.

6. *Захаров В.П., Богданова С.Ю.* Корпусная лингвистика: учебник. - СПб.: Изд-во Санкт-Петербург. гос. ун-та, 2020. - 234 с.

7. *NooJ* [Electronic resource]. URL: <https://nooj.univ-fcomte.fr/> (date of access: 29.05.2024).