Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа
«Цифровые методы в гуманитарных науках»

КУРСОВАЯ РАБОТА

На тему «Смык і Янка: Корпус белорусской поэзии XVIII—XX вв.»
A corpus of Belarusian poetry of the 18th–20th century

Студентка 1 курса
группы № МЦМГН201
Немкович Екатерина Николаевна

Научный руководитель
Ляшевская Ольга Николаевна
Профессор, кандидат филологических наук

Москва, 2021 г.

1. Введение ... 3
2. Проект корпуса белорусской поэзии ... 4
2.1. Существующие поэтические (под)корпусы ... 4
2.2. Требования к разметке корпуса ... 7
2.3. Материал корпуса ... 9
3. Сбор и предварительная обработка текстов ... 11
3.1. Отбор авторов в первую версию ... 11
3.2. Процедура сбора ... 11
3.3. Состав собранного корпуса ... 13
3.4. Количественные характеристики материала первой версии ... 13
4. Заключение ... 16
Литература ... 18
Приложение 1. Пример разметки ... 21
Приложение 2. Список авторов ... 24
Приложение 3. Пример метаданных произведения ... 33
Приложение 4. Информация о репозитории проекта ... 35

1. Введение

Данная работа посвящена процессу создания проекта корпуса белорусской поэзии XVIII—XX вв., а также первому этапу сбора материала указанного корпуса. Необходимость данного проекта обусловлена отсутствием общедоступных корпусов белорусской литературы в целом и поэзии в частности. Наличие такого корпуса позволит проводить исследования белорусской поэзии в русле корпусной лингвистики, стиховедения, цифровых гуманитарных наук и других областей.

К целям проекта относится создание корпуса белорусской поэзии упомянутого периода со стиховедческой, лингвистической и библиографической информацией о произведениях. Задачами работы являются отбор авторов, разработка схемы разметки с необходимыми характеристиками, автоматический сбор текстов и их обработка в соответствии с оговоренными требованиями к разметке. В качестве языка разметки используется XML, частной реализацией которого является схема TEI. Этот стандарт и послужит основой для схемы разметки обсуждаемого корпуса. Рабочий язык программирования данной работы — Python.

В данной работе приводится описание разработанного проекта корпусной разметки и результатов уже завершенного этапа сбора первой когорты авторов посредством веб-скрейпинга, однако создание корпуса в соответствии с поставленными целями потребует проведения значительно большего объема работ, чем обусловлено отдельной курсовой работой. Раздел 2 работы, в котором приводится описание материала корпуса и схема разметки, может служить первичной документацией для дальнейшего наполнения и структуризации корпуса. В разделе 3 обсуждаются прикладные задачи, необходимые для сбора текстового материала, поэтому он может быть также полезен как пример описания алгоритма начального этапа создания корпуса. Данная проблематика, на взгляд автора, недостаточно подробно освещена в учебных пособиях и другой литературе по корпусной лингвистике, которые обычно фокусируются на точке зрения лингвиста и рядового пользователя корпуса. В заключении также обсуждаются некоторые проблемы полученного материала и используемых методов обработки текста, а также дальнейшие направления работы.

2. Проект корпуса белорусской поэзии

2.1. Существующие поэтические (под)корпусы

Поэтический корпус является специальным корпусом, посвященным исследованию поэзии (Захаров, Богданова 2013). На данный момент уже существует ряд поэтических корпусов, которые, как правило, являются частью, или подкорпусом, более общих корпусов. Ниже приведен краткий обзор существующих решений с описанием технических характеристик и разметки поэтических произведений, если информация о них общедоступна.

Британский национальный корпус[1] размечен по стандарту TEI-XML для работы через систему индексации Xaira, но также поддерживает и другие средства обработки. Разметка стихотворений предусматривает деление на строфы, строки, предложения (сегменты, <s>) и слова[2]. Предложения в разметке нумеруются. С помощью POS-теггера CLAWS[3] слова снабжаются лингвистической разметкой, которая включает часть речи, форму и лемму. Стиховедческая разметка отсутствует. Что касается содержимого, корпус включает устные и письменные тексты второй половины XX в. Для того, чтобы тексты были сопоставимыми по размеру, они были обрезаны до нужного объема. Целиком в корпусе приводятся только короткие тексты.

Корпус современного американского английского языка[4] содержит поэзию, однако скорее нацелен на исследования в области лексики, стилистики и тематического моделирования, нежели стиховедения. Художественная литература составляет относительно небольшую долю корпуса и представлена в чуть меньшем объеме, что и газеты, журналы, субтитры к фильмам и ТВ, блоги, академические публикации и веб-источники (по отдельности). К его функциям относятся поиск по словам, частотные словари, определение конкордансов и коллокатов и другие

[1] British National Corpus (http://www.natcorp.ox.ac.uk/)
[2] Poetry. BNC User Reference Guide (http://www.natcorp.ox.ac.uk/docs/URG/cdifwr.html#cdif53)
[3] CLAWS part-of-speech tagger for English (http://ucrel.lancs.ac.uk/claws/)
[4] Corpus of Contemporary American English (https://www.english-corpora.org/coca/)

средства корпусного анализа[5]. Данный корпус обладает инструментами фильтрации по жанрам и стилям. Таким образом, этот корпус хорошо подходит для лингвистических исследований прессы и медиа.

Чешский национальный корпус[6] наряду с НКРЯ является одним из крупнейших славянских корпусов. Его также можно отнести к наиболее разработанным в плане функциональности корпусам — ЧНК включает большое количество специализированных модулей и интерфейсов. Главными направлениями его использования являются получение морфосинтаксической информации о словах, генерация конкордансов и частотных списков и другие виды лингвистического анализа. Примечательной чертой корпуса является наличие географической разметки текстов и возможность исследования использования лексики с помощью карты Чехии. В отдельный модуль AkaLex вынесен интерфейс для работы с чешской академической лексикой.

Национальный корпус русского языка[7], в отличие от ранее упоминавшихся ресурсов, имеет отдельный поэтический подкорпус и специальные средства поиска по нему. В корпус включены два периода: ранний (с XVIII века до середины XX) и современный (середина XX — начало XXI). Тексты снабжены морфологической и лексико-семантической разметкой. Помимо поэтического подкорпуса есть другие подкорпусы с глубокой специализированной разметкой, которая соответствует их задачам. Пропорция художественных произведений в корпусе довольно велика на фоне других примеров — она составляет до 40%[8]. Для стихотворений, помимо обычной для всего корпуса металингвистической информации, указывается метр, размер, длина строки, тип клаузулы, метрическая формула, тип рифмы, количество строф. Очевидно огромное преимущество русского корпуса для исследования стихотворной речи. Однако разметка поэтического подкорпуса имеет и определенные недостатки. В частности, помещение стиховедческой информации

[5] COCA Overview (https://www.english-corpora.org/coca/help/coca2020_overview.pdf)
[6] Czech National Corpus (https://www.korpus.cz/)
[7] Национальный корпус русского языка (https://ruscorpora.ru/)
[8] Состав и структура корпуса. Национальный корпус русского языка (https://ruscorpora.ru/new/corpora-structure.html)

вместе с металингвистической не позволяет указывать характеристики отдельных элементов произведения. Например, метр может меняться или чередоваться на протяжении стихотворения, и присвоение единой ритмической формулы целому произведению приводит к потере информации о некоторых составных элементах. Кроме того, реальный ритм не всегда соответствует «идеальной» метрической формуле.

Башкирский поэтический корпус[9] также предоставляет возможности работы со стиховедческими характеристиками текстов. Помимо лексического и грамматического поиска в нем доступны поиск по произведениям с выбранным метром и поиск по рифме, то есть тексты снабжены стиховедческой разметкой. Это составляет большое преимущество данного ресурса. К функциональным недостаткам интерфейса можно отнести тот факт, что поисковая выдача включает только строку со словоформой и соседние с ней.

Русско-французский поэтический корпус[10] является примером параллельного корпуса, который позволяет установить влияние одной литературной традиции на другую. В его функциональность входит поиск по словоформам и леммам на русском и французском языках с параллельным просмотром франко- и русскоязычных версий текста, который отвечает критериям поиска. Стиховая разметка представлена в виде выделения зоны рифмовки как локации искомого слова. Поскольку проект нацелен скорее на исследование лексики и синтаксиса, в разметке отсутствует информация о ритме и других стиховедческих характеристиках.

Отдельного упоминания заслуживает разрабатываемая Е. Казарцевым и В. Вашченковым универсальная компьютерная система анализа ритмики текстов на разных языках[11][12] (Казарцев 2017). Она совмещает корпусный подход с глубоким автоматическим анализом метра и ритма текста. Система предоставляет возможность автоматической акцентуации, определяет метр и размер, осуществляет

[9] Башкирский поэтический корпус (http://web-corpora.net/bashcorpus/search/index.php)
[10] Русско-французский поэтический корпус (http://www.nevmenandr.net/fr/index.php)
[11] Corpus – Marked Texts (http://vikvas1r.beget.tech/)
[12] Данная разработка осуществлялась в рамках проекта, поддержанного Российским научным фондом (грант № 16-18-10250).

статистический учет использования метров, визуализирует метрический профиль. Хотя данное веб-приложение еще находится в разработке и адаптировано под русский язык, оно уже используется для работы с белорусской поэзией — Т. Земскова при его помощи провела исследование эволюции белорусского 4-стопного ямба (Земскова, Казарцев).

2.2. Требования к разметке корпуса

Разметка обсуждаемого в данной работе корпуса поэзии подразумевает три основных аспекта: библиографический (в данном случае он совпадает с металингвистическим), лингвистический и стиховедческий. В качестве языка разметки используется TEI-XML[13]. Для целей автоматической обработки и анализа также подразумевается хранение упрощенной версии текстов с приведением к нижнему регистру и без знаков препинания.

Металингвистическая разметка включает сведения об авторе и произведении в целом: имя автора, название, дату создания и первой публикации, язык оригинала и имя автора перевода для переводных произведений, жанр, издание, а также сведения об онлайн-источнике и ссылку на него. Все эти характеристики помещаются в <TEI header>.

Лингвистическая разметка заключается в традиционных для корпусов токенизации и POS-тегировании. Каждый токен снабжается морфосинтаксической информацией в соответствии со стандартом POS-тегов Universal Dependencies[14]:

```
<w pos="VERB" lemma="быць" msd="Past:Imp:Act:Masc:Sing">быў</w>
```

Указывается часть речи, лемма и грамматическая информация о слове. Данная разметка хранится в контейнере <ana type="linguistic">, который относится к отдельной строке.

Стиховедческая разметка характеризуется наибольшей сложностью как в теоретическом анализе, так и в реализации, поскольку ритм проявляется на всех структурных уровнях от отдельной фонемы до строфики (Лотман 1996).

---

[13] Verse. Text Encoding Initiative (https://tei-c.org/release/doc/tei-p5-doc/en/html/VE.html#VEME)
[14] UD for Belarusian (https://universaldependencies.org/be/)

Для отдельной строки размечается ее номер, фактическая ритмическая схема, метр (в случае, если он регулярен) и количество стоп:

```
<l n="1" met="-+|-+|-+|--|-+|-/" metre="iambus" feet="5">
```

Ударные слоги отмечаются знаком плюса, а безударные — минуса. Знак черты ‘|’ используется для обозначения границы стопы. Кроме того, в контейнере <ana type="rhyme"> указывается идентификатор рифмы в стихотворении в виде латинской буквы, а также ее тип (мужская, женская, дактилическая). В контейнере <ana type="metre"> хранится версия стиха с сегментацией по стопам и слогам с указанием их ударности.

Стихи группируются в строфу и снабжаются номерами, указаниями на тип строфы по количеству строк и схему рифмовки, которая генерируются на основе идентификаторов отдельных строк:

```
<lg n="1" type="quatrain" rhyme="AbbA">
```

Для рифмы используется традиционное обозначение латинскими буквами разного регистра. Женские рифмы обозначаются заглавными буквами, а мужские — строчными. Строфы, которые разделены графически, но рифмуются между собой, дополнительно группируются при помощи контейнера <lg>.

Самый высокий уровень стиховедческой разметки передает характеристики, которые относятся к тексту целиком. Они помещаются в контейнер <formal>:

```
<formal rhyme="AbbA|AbbA|ccD|eDe" metre="iambus"/>
```

Параметр ‘rhyme’ иллюстрирует строфику и схему рифмовки стихотворения, а ‘metre’ — метр произведения (в случае, если текст имеет ритмически регулярную структуру). Стоит обратить внимание на то, что полное кодирование схемы рифмовки произведения имеет смысл только для коротких стихотворных форм, в анализе который оно и составляют бóльшую аналитическую ценность. Для крупных произведений, рифмы которых нельзя полностью разметить при помощи латинского алфавита, данная форма кодирования должна производиться в пределах одной строфы.

В случаях, когда четкая классификация или членение не представляется возможным, соответствующий тип разметки не генерируется или генерируется только с теми значениями параметров, которые доступны.

Подробнее с примером разметки стихотворения можно ознакомиться в Приложении 1.

2.3. Материал корпуса

Временными границами для поиска материалов корпуса служат начало XVIII века и конец XX. Если малое количество произведений XVIII века избавляет разработчика корпуса от мучительного выбора среди множества доступных вариантов, то для верхней границы необходима бóльшая точность. Поэтому в рамках данной работы «конец XX века» уточняется как конец советской эпохи.

Помимо оригинальных произведений в корпус включаются также и переводные, поскольку на определенных этапах развития белорусская литература не была исключительно белорусскоязычной. Поэты начала XVIII века писали на латыни, а такие авторы начала XIX века, как Я. Барщевский и В. Сырокомля, традиционно включаются в белорусский литературный канон, хотя и в переводе с польского.

Список авторов текущей версии корпуса базируется на антологии белорусской поэзии, изданной в 1993 году под редакцией Р. Бородулина (Барадулін 1993). Данное издание сейчас является последней опубликованной антологией белорусской поэзии, и его верхняя хронологическая граница (год издания) совпадает с рамками корпуса. Она включает и произведения, изданные ранее XVIII века, поэтому часть авторов не попала в итоговый перечень. После отсева по хронологическому критерию в списке осталось 293 автора. Этот список был дополнен биографической информацией (годы жизни, пол), а также ссылками на страницы поэтов на Википедии и их профили на электронных библиотеках, посвященных белорусской литературе. Каждому автору был присвоен идентификатор в виде порядкового номера. Результат лег в основу CSV-файла, который использовался для последующего сбора текстов из онлайн-источников. Список можно увидеть в Приложении 2.

Для создания прототипа корпуса лучше всего подходят онлайн-ресурсы, поскольку обработка их содержимого требует наименьших затрат труда и времени, поэтому логичным началом сбора материала было именно обращение к ним. К общедоступным Интернет-ресурсам, на которых творчество белорусских поэтов представлено наиболее полно, относятся «Беларуская палічка»[15] (далее knihi.com) (72) и Вершы.ru[16]. После сравнения количества страниц поэтов из списка на данных веб-сайтах было установлено, что «Беларуская палічка» лучше подходит в качестве ресурса для первичного сбора материала (73 поэта против 60 на Вершы.ру). Помимо количества представленных поэтов «Беларуская палічка» существенно превосходит «Вершы» по количеству произведений. Кроме того, knihi.com обладает достаточно простой структурой страниц, а тексты на нем часто снабжены крайне полезными метаданными. Тем не менее, «Вершы» также могут впоследствии для обогащения корпуса текстами, которые не входят в библиотеку knihi.com.

[15] Беларуская палічка. Беларуская электронная бібліятэка (https://knihi.com/)
[16] Вершы.ru. Нацыянальны паэтычны партал (http://vershy.ru/)

3. Сбор и предварительная обработка текстов

3.1. Отбор авторов в первую версию

Чтобы ограничить объем текстов, которое подлежат сбору и обработке в рамках первой версии корпуса, из авторов, входящих в вышеописанный список, были выбраны 79 представителей. Критериями выбора были важность автора для канона и приблизительная равномерность распределения авторов во времени. Первому критерию удовлетворяли авторы, которые относятся к Народным поэтам Беларуси и стабильно входят в школьный и университетский канон преподавания, например, в учебные программы (Губская 2016). После добавления таких авторов список также по возможности балансировался хронологически, чтобы в корпусе были репрезентированы поэты разных поколений. Однако важно упомянуть, что первую когорту авторов этого проекта нельзя рассматривать как некую объективную категорию, и они были выбраны для сокращения объема работ по пилотной версии и тестирования выбранных методов.

После предварительного отбора авторы были проверены на наличие профиля на knihi.com. В результате отсеяно 6 авторов, которые не были представлены на ресурсе. В итоговый список вошло 73 автора, ссылки на страницы которых были добавлены в таблицу для последующего веб-скрейпинга. Полный список с соответствующими отметками можно увидеть в Приложении 2.

3.2. Процедура сбора

Процесс сбора текстов включает нескольких этапов: сбор ссылок на страницы произведений, их организация, сбор текстов произведений в виде HTML и парсинг их метаданных. Для этих целей были использованы скрипты на Python с подключением библиотеки Beautiful Soup[17] (далее BS4).

Сбор ссылок на страницы произведений производится при помощи скрипта для парсинга, адаптированного конкретно под используемый ресурс. Грамотное

[17] Beautiful Soup Documentation (https://www.crummy.com/software/BeautifulSoup/bs4/doc/)

обнаружение и категоризация ссылок имеют большое значение для оптимизации алгоритма сбора, поскольку на сайте поддерживаются форматы произведений, которые не могут быть использованы в рамках данного проекта. Помимо текстового формата, произведения на knihi.com хранятся также в форматах PDF, DJVU, MP3, MP4. Кроме того, список произведений на странице автора часто содержит сразу две или более ссылки на один и тот же текст. Как правило, у произведения есть прямая ссылка на файл и также ссылка на веб-страницу. Например, если это текстовый файл, ресурс дает пользователю возможность скачать файл EPUB или читать произведение на странице сайта. В некоторых случаях при данных ссылках также встречаются ссылки на страницы других авторов (например, если издание переводное или коллективное). Таким образом, механический перебор ссылок имеет низкую продуктивность. Для решения вышеупомянутой проблемы страница автора обрабатывается BS4, а каждый элемент списка произведений делится на составные части при помощи регулярных выражений. В CSV-таблицу, которая впоследствии еще будет дополнена другими метаданными, сохраняется название произведения, формат, все доступные ссылки и их подписи. Для каждого произведения генерируется идентификатор вида ‘ID автора_порядковый номер’.

Скачивание текстовых файлов в формате HTML происходит при помощи скрипта, который проходит циклом по сгенерированной ранее таблице, выбирая только ссылки на произведения в текстовом формате. На диск сохраняется не страница целиком, а содержимое контейнера с основным текстом. Навигационные элементы сайта игнорируются. Скрипт сохраняет файлы под названием вида ‘ID автора_порядковый номер’ в папки, названия которых включают идентификатор и имя автора.

Парсинг страниц произведений позволяет существенно обогатить метаданные, так как структура страницы на knihi.com предусматривает наличие JavaScript-комментариев, в которые заключена такая библиографическая информация, как издание, годы создания и первой публикации, жанр, язык оригинала и автор перевода (для переводных произведений). Иногда указываются подзаголовки и другие пометки автора, издателя или редактора сайта. Далеко не все тексты

снабжаются такими данными, однако их доля велика. Всего при парсинге было обнаружено 31 параметра метаданных, но стабильно используется примерно 5–10 из них. Очевидно, что некоторые из них были техническими или тестовыми данными, у отдельных параметров есть дубликаты. Вне зависимости от качества описанных метаданных содержимое тегов добавляется в таблицу с данными о текстах в соответствующие столбцы для последующей обработки.

3.3. Состав собранного корпуса

Агрегация метаданных списка произведений позволит ближе познакомиться с собранным материалом. Однако перед этим знакомством важно упомянуть, что описываемая в данном разделе совокупность текстов пока является **не поэтическим корпусом, а корпусом произведений поэтов**. Как упоминалось выше, сбор текстов производится по списку авторов, а жанр отдельных произведений невозможно определить до загрузки текстов с метаданными. Многие авторы поэтических произведений писали также и прозу. Также существуют произведения, в которых сочетаются элементы прозы и поэзии, например, «Паўлінка» Я. Купалы. Более того, среди работ выбранных авторов иногда встречаются и нехудожественные тексты.

На данный момент отделение поэтических, прозаических и смешанных произведений друг от друга является задачей, которую только предстоит решить. Однако это может впоследствии стать и преимуществом корпуса — уже разработанные скрипты сбора текстов с таким же успехом могут работать и с творчеством прозаиков. Таким образом, в будущем корпус способен включить в себя и прозаические тексты без серьезных изменений общей архитектуры. Главным на данный момент препятствием является отсутствие схем разметки прозаических и смешанных текстов, разработка которых не входит в цели данной работы и заслуживает отдельных исследований.

3.4. Количественные характеристики материала первой версии

С помощью указанных выше операций было собрано 5948 текстов. Наиболее богато представлены белорусские классики. Более 100 произведений есть у

следующих 12 авторов: Я. Колас (1052 произведений), Я. Купала (904), У. Караткевіч (443), П. Макаль (436), Я. Сіпакоў (433), А. Звонак (342), П. Панчанка (305), М. Лужанін (228), К. Крапіва (222), А. Гарун (177), У. Жылка (116), Ц. Гартны (109). Первые 15 авторов по количеству произведений можно увидеть на Рис. 1.

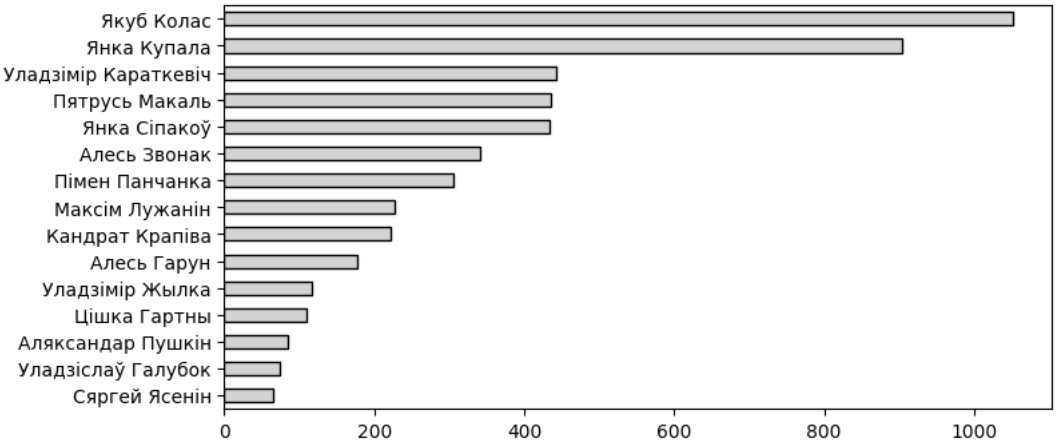

Рис. 1. Количество произведений отдельных авторов.

Среди первых 25 авторов по количеству текстов также присутствуют представители других восточнославянских литератур: А. Пушкин (84), С. Есенин (65 текстов, 15-й), И. Крылов (48, 18-й), Т. Шевченко (44, 19-й), М. Лермонтов (42, 21-й). Последние, конечно же, не входят в список авторов корпуса и попали в него посредством переводов. Это указывает на активную переводческую деятельность многих поэтов — всего в корпусе обнаружен 31 переводчик. Среди оригиналов переводных произведений подавляющее большинство составляют русские (283), далее следуют польские (68), украинские (54) и французские (24). Пропорцию этих языков можно оценить на Рис. 2. Каждый из остальных языков используется менее, чем в 10 произведениях.

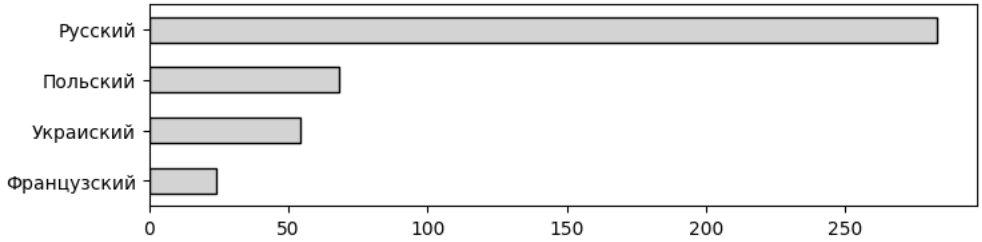

Рис. 2. Языки оригиналов переводных произведений.

Параметр ‘StyleGenre’, который среди метаданных является основным индикатором для определения жанров и разделения текстов на прозу и поэзию, указан далеко не у всех произведений. Данный параметр остался пустым у 2012 из 5948 текстов. Таким образом, без дополнительной обработки жанр можно определить у двух третей текстов, или 66%. Как стихотворения помечены 2975 текстов, как поэмы — 92.

Отдельной проверке следует подвергнуть такие смешанные значения, как ‘мастацкі/верш,мастацкі/казка’. Однако это не является приоритетной задачей, поскольку такие пометки используются достаточно редко.

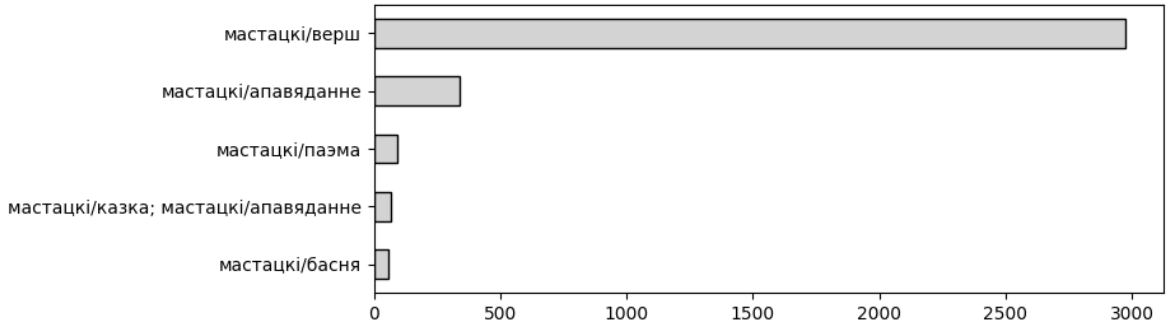

Рис. 3. Самые распространенные значения параметра ‘StyleGenre’.

4. Заключение

В ходе работы над описываемым в данном тексте проектом были разработаны требования к разметке корпуса, составлен рабочий список белорусских поэтов, собраны тексты первой когорты авторов и доступные для них их метаданные. Собранные произведения представлены 5948 текстами, которые были снабжены уникальными идентификаторами и организованы по авторам (информацию о репозитории проекта можно найти в Приложении 4). Упомянутые результаты можно назвать удовлетворительными, однако у схемы разметки, алгоритмов сбора и парсинга, а также у самого материала корпуса до сих пор остались нерешенные проблемы.

Создание схемы разметки представляется собой один из самых сложных этапов проекта, поскольку у стандарта TEI-XML по сей день нет достаточного количества примеров с множественными уровнями разметки. Как правило, корпусы имеют только лингвистическую разметку, реже — отдельные элементы стиховой. За недостатком примеров автору схемы приходится принимать решения о хранении информации на уровнях с совершенно разными масштабами — от слога до строфы. Более того, некоторые из структурных уровней могут снабжаться одновременно несколькими типами разметки. Помимо имеющихся параметров, схема разметки должна, как минимум, быть дополнена обозначениями метрических формул и других стиховых характеристик, которые соответствуют традиции восточнославянского (в частности, русского) литературоведения, поскольку в стандарте TEI она не представлена. При этом некоторые из обозначений стандарта TEI-XML все предоставляют информацию, которая, например, не предусмотрена разметкой НКРЯ (к такой информации относится ритмическая схема одиночного стиха). Для максимальной полноты описания эти две традиции должны сосуществовать в одной схеме. Кроме того, необходимо решить проблему несоответствия графической и ритмической структуры стихотворения, поскольку анализ исключительно графической версии часто не позволяет корректно определить метр и размер произведения. Важными траекториями дальнейшей разработки схемы могут стать

добавление фонетической разметки и возможность хранения нескольких редакций одного текста (Пильщиков 2017).

Что касается наполнения, полученный корпус нельзя назвать репрезентативным. Это иллюстрируется тем фактом, что некоторые народные поэты Беларуси представлены в нем в меньшем объеме, чем Пушкин или Есенин. Нужно также принимать во внимание, что издания, по которым приводятся тексты, не всегда отвечают филологическим и текстологическим стандартам качества. Однако целью данной работы является создание проекта и технической базы корпуса, а не оптимальное его наполнение. Корректировка полноты и репрезентативности может быть произведена после его завершения.

Некоторые трудности возникли с классификацией произведений по жанрам и их внутренней структурой. Поскольку относящиеся к жанру метаданные не заполнены у трети собранных произведений, без дополнительной обработки невозможно автоматически установить, являются ли они поэзией. Этот вопрос может быть решен с использованием таких методов компьютерной классификации, как логистическая регрессия. Также предстоит решить проблему определения вложенности разделов для крупных произведений и сборников, которые приводятся одним файлом.

Литература

Анталогія беларускай паэзіі у 3 т. Пад рэд. Р. Барадуліна. Мінск: Мастацкая літаратура, 1993.

Гаспаров М. Л. Русский стих начала XX века в комментариях. М.: Фортуна Лимитед, 2001. С. 112–139.

Гаспаров М. Л. Очерк истории европейского стиха. М.: Фортуна Лимитед, 2003. С. 169–217.

Гаспаров М.Л. Современный русский стих: метрика и ритмика. М.: Наука, 1974.

Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 71–113.

Губская В.М. Беларуская літаратура. Вучэбная праграма ўстановы вышэйшай адукацыі, 2016 (https://elib.bsu.by/handle/123456789/186934)

Жирмунский В. М. Теория стиха. Л.: Советский писатель, 1975. С. 26–162.

Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». СПб.: СПбГУ. РИО. Филологический факультет, 2013.

Земскова Т. А., Казарцев Е. В. Становление и эволюция белорусского 4-стопного ямба // Russian Literature. 2021. С. 1–17.

Казарцев Е. В. Сравнительное стиховедение: метрика и ритмика. СПб.: Издательство РГПУ им. А. И. Герцена, 2017. С. 21–37.

Лотман Ю. М. Анализ поэтического текста: Структура стиха // О поэтах и поэзии. СПб., 1996. С. 18–252.

Орехов Б. В. Башкирский стих XX века. Корпусное исследование. СПб.: Алетейя, 2019.

Орехов Б. В., Савчук С. О. Акцентологический корпус как инструмент для исследования русского ударения // Труды Института русского языка им. В. В. Виноградова (№21). М.: Институт русского языка им. В. В. Виноградова РАН, 2019. С. 61–82.

Пильщиков И. А. О задачах поэтических корпусов // Труды Института русского языка им. В. В. Виноградова (№11). М.: Институт русского языка им. В.В. Виноградова РАН, 2017. С. 332–337.

Пильщиков И. А. Понятия «стих», «метр» и «ритм» в русской стиховедческой традиции // Труды Института русского языка им. В. В. Виноградова (№11). М.: Институт русского языка им. В. В. Виноградова РАН, 2017. С. 12–30.

Пильщиков И. А., Старостин А. С. Основные проблемы автоматизации базовых процедур ритмико-синтаксического анализа силлабо-тонических текстов // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 298–316.

Прикладная и компьютерная лингвистика. Под ред.: Николаев И. С., Митренина О. В., Ландо Т. М. М.: Издательская группа URSS, 2017.

Резникова Т. И. Славянская корпусная лингвистика: современное состояние ресурсов // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 402–461.

Тарановский К. Основные задачи статистического изучения славянского стиха // О поэзии и поэтике. М.: Языки русской культуры, 2000. С. 234–256.

Холшевников В. Е. Основы стиховедения: Русское стихосложение. СПб.: Филологический факультет СПбГУ; М.: Издательский центр «Академия», 2004. С. 3–84.

Plamondon, M. R. Virtual verse analysis: Analysing patterns in poetry. Literary and Linguistic Computing, 21 (Suppl. 1), 2006. Pp. 127–141.

Интернет-ресурсы

Башкирский поэтический корпус (http://web-corpora.net/bashcorpus/search/index.php)

Национальный корпус русского языка (https://ruscorpora.ru/)

Русско-французский поэтический корпус (http://www.nevmenandr.net/fr/index.php)

Состав и структура корпуса. Национальный корпус русского языка (https://ruscorpora.ru/new/corpora-structure.html)

British National Corpus (http://www.natcorp.ox.ac.uk/)

CLAWS part-of-speech tagger for English // UCREG (http://ucrel.lancs.ac.uk/claws/)

COCA Overview // Corpus of Contemporary American English (https://www.english-corpora.org/coca/help/coca2020_overview.pdf)

Corpus – Marked Texts (http://vikvas1r.beget.tech/)

Corpus of Contemporary American English (https://www.english-corpora.org/coca/)

Czech National Corpus (https://www.korpus.cz/)

Poetry // British National Corpus User Reference Guide (http://www.natcorp.ox.ac.uk/docs/URG/cdifwr.html#cdif53)

UD for Belarusian // Universal Dependencies (https://universaldependencies.org/be/)

Verse // Text Encoding Initiative (https://tei-c.org/release/doc/tei-p5-doc/en/html/VE.html#VEME)

Приложение 1. Пример разметки

```
<TEI xmlns='http://www.tei-c.org/ns/1.0'>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type='main'>Санет ("Замёрзла ноччу шпаркая
крыніца...")</title>
        <author>
          <persName>
            <forename>Максім</forename>
            <surname>Багдановіч</surname>
          </persName>
        </author>
      </titleStmt>
      <sourceDesc>
        <bibl type='onlineSource'>
          <name>БЕЛАРУСКАЯ ПАЛІЧКА</name>
          <idno
type='url'>https://knihi.com/Maksim_Bahdanovic/Saniet_Zamiorzla_noccu_spa
rkaja_krynica.html</idno>
          <bibl type='originalSource'>
            <title>Лазарук, М.А. Беларуская літаратура: вучэб. дапам. для
8-га кл. устаноў агульнай сярэдняй адукацыі з беларус. і рус. мовамі
навучання / М.А.Лазарук, В.І.Русілка, І.М.Слесарава. – Мінск: Нац. ін-т
адукацыі, 2011.</title>
            <date type='written'>1912</date>
            <date type='print'>1912</date>
          </bibl>
        </bibl>
      </sourceDesc>
      <profileDesc>
        <textClass>
          <form type='poetry'>Паэзія</form>
          <genre type='poem'>Верш</genre>
          <genre type='sonnet'>Санет</genre>
        </textClass>
      </profileDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <front>
      <head>Санет ("Замёрзла ноччу шпаркая крыніца...")</head>
    </front>
    <body>
      <formal rhyme='AbbA|AbbA|ccD|eDe' metre='iambus'/>
      <lg n='1' type='quatrain' rhyme='AbbA'>
        <l n='1' met='-+|-+|-+|--|-+|-/' metre='iambus' feet='5'>
          Замёрзла ноччу шпаркая крыніца;

          <ana type='linguistic'>
            <w pos='VERB' lemma='мерзнуць'
msd='Past:Perf:Act:Fem:Sing'>замёрзла</w>
            <w pos='ADV' lemma='ноччу'>ноччу</w>
```

```
            <w pos='ADJ' lemma='шпаркая' msd='Nom:Fem:Sing'>шпаркая</w>
            <w pos='NOUN' lemma='крыніца'
msd='Nom:Fem:Sing:Inan'>крыніца</w>
            <pc pos='PUNCT' lemma=';'>;</pc>
          </ana>

          <ana type='rhyme'>
            Замёрзла ноччу шпаркая крыніца<rhyme label='A'
type='fem'>крыніца</rhyme>;
          </ana>

          <ana type='rhythm'>
            <seg type='foot'>
              <seg type='syll'>За</seg>
              <seg type='syll' stress='ictus'>мёрз</seg>
            </seg>
            <seg type='foot'>
              <seg type='syll'>ла</seg>
              <seg type='syll' stress='ictus'>ноч</seg>
            </seg>
            <seg type='foot'>
              <seg type='syll'>чу</seg>
              <seg type='syll' stress='ictus'>шпар</seg>
            </seg>
            <seg type='pyrrhic'>
              <seg type='syll'>ка</seg>
              <seg type='syll'>я</seg>
            </seg>
            <seg type='foot'>
              <seg type='syll'>кры</seg>
              <seg type='syll' stress='ictus'>ні</seg>
            </seg>
            <seg type='clausula'>
              <seg type='syll'>ца</seg>
            </seg>
          </ana>
        </l>
        <l n='2' met='-+|-+|-+|--|-+/'>Твая пара, зімовая нуда!</l>
        <l n='3' met='-+|-+|-+|--|-+/'>Цяпер няма ўжо руху ні сляда,</l>
        <l n='4' met='-+|-+|-+|-+|-+|-/'>І нават зверху слоем снег
лажыцца.</l>
      </lg>

      <lg n='2' type='quatrain' rhyme='AbbA'>
        <l n='5' met='-+|-+|-+|-+|-/'>Ды ўсё дармо, бо там, пад ім,
струіцца</l>
        <l n='6' met='-+|--|-+|--|-+/'>Магутная, жывучая вада.</l>
        <l n='7' met='-+|-+|-+|--|-+/'>Чакай! Яе йшчэ прыйдзе чарада!</l>
        <l n='8' met='-+|-+|-+|-+|-+|-+|-/'>Здалее хваляў хор на вольны
свет прабіцца.</l>
      </lg>

      <lg n='3' type='stanza' rhyme='ccDeDe'>
        <lg type='tercet' rhyme='ccD'>
```

```
          <l n='9' met='-+|-+|-+|--|-+/'>Прыклаў я гэты сімвал да
сябе,</l>
          <l n='10' met='-+|--|-+|--|-+/'>Схіліўшыся ў надсільнай
барацьбе,</l>
          <l n='11' met='--|-+|-+|-+|-+|-/'>І разгадаў прыроды роднай
словы.</l>
        </lg>

        <lg n='4' type='tercet' rhyme='eDe'>
          <l n='12' met='+-|-+|-+|-+|-+/'>Як — прамаўчу, бо кожны з вас —
паэт.</l>
          <l n='13' met='-+|-+|-+|-+|-+|-/'>Рассейце ж самі лёгкі змрок
прамовы,</l>
          <l n='14' met='-+|-+|-+|-+|-+/'>Сваёй душы туды праліце
свет!</l>
        </lg>
      </lg>
    </body>
  </text>
</TEI>
```

Приложение 2. Список авторов

ID	*Автор*	*Включен(а) в 1-ю когорту*	*Наличие на knihi.com*
1	Іаахім Храптовіч		
2	Міхал Карыцкі	Да	
3	Ян Баршчэўскі	Да	Есть
4	Паўлюк Багрым	Да	Есть
5	Францішак Савіч		
6	Ігнат Легатовіч		
7	Ян Чачот	Да	Есть
8	Адэля з Устроні		
9	Вінцэнт Дунін-Марцінкевіч	Да	
10	Арцём Вярыга-Дарэўскі		
11	Ялегі Пранціш Вуль		
12	Аляксандр Рыпінскі		
13	Уладзіслаў Сыракомля	Да	Есть
14	Вінцэсь Каратынскі		
15	Кастусь Каліноўскі	Да	Есть
16	Вайніслаў Савіч-Заблоцкі		
17	Аляксандр Ельскі		
18	Францішак Багушэвіч	Да	Есть
19	Янка Лучына	Да	Есть
20	Фелікс Тапчэўскі		
21	Адам Гурыновіч	Да	Есть
22	Альгерд Абуховіч		
23	Лявон Вітан-Дубейкаўскі		
24	Бруно Каратынскі		
25	Карусь Каганец	Да	Есть
26	Цётка	Да	Есть
27	Дзядзька Пранук		
28	Янка Купала	Да	Есть
29	Якуб Колас	Да	Есть
30	Эдзюк Будзька		
31	Максім Багдановіч	Да	Есть
32	Алесь Гарун	Да	Есть

33	Альберт Паўловіч		
34	Дзед Доніс		
35	Гальяш Леўчык		
36	Уладзіслаў Галубок	Да	Есть
37	Вацлаў Ластоўскі		
38	Юзя Шчупак		
39	Цішка Гартны	Да	Есть
40	Алесь Гурло		
41	Анатоль Дзяркач		
42	Андрэй Зязюля		
43	Янук Д.		
44	М. Арол		
45	Янка Журба	Да	Есть
46	Канстанцыя Буйло	Да	Есть
47	І. Піліпаў		
48	Фабіян Шантыр		
49	Змітрок Бядуля	Да	
50	Хведар Чарнышэвіч	Да	
51	Лявон Лобік		
52	Альфонс Петрашкевіч		
53	Кандрат Лейка		
54	Зоська Верас	Да	Есть
55	Леапольд Родзевіч		
56	Юзюк Фарботка		
57	Казімір Сваяк		
58	Міхась Ганчарык		
59	Міхайла Грамыка		
60	Міхась Чарот	Да	Есть
61	Кандрат Крапіва	Да	Есть
62	Ігнат Канчэўскі		
63	Макар Краўцоў		
64	Алесь Смаленец		
65	Міхась Клімковіч		
66	Аркадзь Моркаўка		
67	Уладзімір Жылка	Да	Есть

68	Уладзімір Дубоўка	Да	Есть
69	Нічыпар Чарнушэвіч		
70	Язэп Пушча	Да	Есть
71	Міхась Машара		
72	Анатоль Вольны		
73	Тодар Кляшторны		
74	Наталля Арсеннева	Да	Есть
75	Піліп Пестрак		
76	Мікола Хведаровіч		
77	Паўлюк Трус		
78	Адам Русак	Да	Есть
79	Алесь Дудар		
80	Уладзімір Хадыка		
81	Алесь Вечар		
82	Юрка Гаўрук		
83	Пятрусь Броўка	Да	Есть
84	Пятро Глебка	Да	Есть
85	Янка Бобрык		
86	Віктар Казлоўскі		
87	Янка Туміловіч		
88	Міхась Васілёк		
89	Ізраіль Плаўнік		
90	Андрэй Александровіч		
91	Сяргей Фамін		
92	Сяргей Новік-Пяюн		
93	Алесь Салагуб		
94	Алесь Звонак	Да	Есть
95	Сяргей Дзяргай	Да	Есть
96	Раман Сабаленка		
97	Станіслаў Шушкевіч		
98	Яўгенія Пфляўмбаўм		
99	Алесь Мілюць		
100	Міхась Багун		
101	Мікола Зосім		
102	Пятрусь Граніт		

103	Сяргей Дарожны		
104	Зінаіда Бандарына		
105	Максім Лужанін	Да	Есть
106	Сяргей Ракіта		
107	Павел Сушко		
108	Сяргей Крывец		
109	Алесь Пруднікаў		
110	Зяма Піваварaў		
111	Змітро Віталін		
112	Ларыса Геніюш	Да	Есть
113	Алесь Жаўрук		
114	Уладзімір Корбан		
115	Змітрок Астапенка		
116	Алесь Дубровіч		
117	Аляксей Зарыцкі	Да	
118	Сцяпан Ліхадзіеўскі		
119	Васіль Вітка	Да	Есть
120	Павел Пруднікаў		
121	Анатоль Астрэйка		
122	Юлій Таўбін		
123	Клім Грыневіч		
124	Анатоль Іверс		
125	Максім Танк	Да	Есть
126	Аляксей Русецкі	Да	Есть
127	Андрэй Ушакоў		
128	Сяргей Грахоўскі	Да	Есть
129	Эдзі Агняцвет	Да	Есть
130	Сяргей Астрэйка		
131	Аркадзь Куляшоў	Да	Есть
132	Валянцін Таўлай		
133	Уладзімір Клішэвіч		
134	Антон Бялевіч		
135	Рыгор Няхай		
136	Рыгор Жалязняк		
137	Мікола Сямашка		

138	Васіль Матэвушаў		
139	Ніна Тарас		
140	Масей Сяднеў		
141	Міхась Калачынскі	Да	Есть
142	Янка Непачаловіч		
143	Пімен Панчанка	Да	Есть
144	Уладзімір Рагуцкі		
145	Мікола Сурначоў		
146	Леанід Гаўрылаў		
147	Алесь Бачыла	Да	Есть
148	Эдуард Валасевіч		
149	Паўлюк Прануза		
150	Міхась Панкрат		
151	Алесь Бажко		
152	Уладзімір Шахавец		
153	Кастусь Кірэенка		
154	Аркадзь Гейнэ		
155	Авар’ян Дзеружынскі		
156	Мікола Аўрамчык		
157	Аляксей Коршак		
158	Пятро Прыходзька		
159	Аркадзь Марціновіч		
160	Аляксей Пысін	Да	Есть
161	Алесь Астапенка		
162	Іван Муравейка		
163	Мікола Гамолка		
164	Алесь Салавей		
165	Анатоль Вялюгін	Да	Есть
166	Віктар Швед		
167	Марк Смагаровіч		
168	Хведар Жычка		
169	Уладзімір Ляпёшкін		
170	Еўдакія Лось	Да	Есть
171	Уладзіслаў Нядзведскі		
172	Алесь Ставер		

173	Валянцін Тарас		
174	Алесь Барскі		
175	Уладзімір Караткевіч	Да	Есть
176	Мікола Арочка	Да	Есть
177	Алег Лойка	Да	Есть
178	Сцяпан Гаўрусеў		
179	Мікола Янчанка		
180	Ніл Гілевіч	Да	Есть
181	Анатоль Вярцінскі	Да	Есть
182	Генадзь Кляўко	Да	Есть
183	Леанід Яўменаў		
184	Пятрусь Макаль	Да	Есть
185	Іван Калеснік		
186	Віктар Шымук		
187	Юрась Свірка	Да	Есть
188	Віктар Хаўратовіч		
189	Віктар Ракаў		
190	Мікола Кусянкоў		
191	Рыгор Барадулін	Да	Есть
192	Уладзімір Лапковіч		
193	Васіль Зуёнак	Да	Есть
194	Іван Летка		
195	Уладзімір Паўлаў		
196	Янка Сіпакоў	Да	Есть
197	Раман Гармола		
198	Міхась Рудкоўскі		
199	Яўген Міклашэўскі		
200	Яўген Крупенька		
201	Генадзь Бураўкін		
202	Яўген Шабан		
203	Пятро Сушко		
204	Алесь Наўроцкі		
205	Міхась Стральцоў	Да	Есть
206	Сымон Блатун		
207	Мікола Купрэеў		

208	Барыс Беляжэнка		
209	Уладзімір Верамейчык		
210	Данута Бічэль-Загнётава	Да	Есть
211	Нэлі Тулупава	Да	Есть
212	Іосіф Скурко		
213	Уладзімір Карызна		
214	Анатоль Грачанікаў	Да	Есть
215	Уладзімір Скарынкін		
216	Хведар Чэрня		
217	Васіль Макарэвіч		
218	Анатоль Канапелька		
219	Зніч		
220	Леанід Дайнека		
221	Ян Чыквін		
222	Ніна Загорская		
223	Васіль Жуковіч		
224	Вячаслаў Дашкевіч		
225	Анатоль Сербантовіч	Да	Есть
226	Іван Ласкоў		
227	Мікола Маляўка		
228	Вера Вярба		
229	Іван Арабейка		
230	Сяргей Панізнік	Да	Есть
231	Рыгор Яўсееў		
232	Мікола Чарняўскі		
233	Мар’ян Дукса		
234	Мікола Федзюковіч		
235	Ніна Мацяш	Да	Есть
236	Генадзь Дзмітрыеў		
237	Казімір Камейша		
238	Уладзімір Лісіцын	Да	Есть
239	Марыя Шаўчонак		
240	Вольга Іпатава		
241	Таіса Бондар		
242	Рыгор Семашкевіч		

243	Уладзімір Дзюба		
244	Валянціна Коўтун		
245	Уладзімір Някляеў	Да	Есть
246	Віктар Гардзей		
247	Алег Салтук		
248	Сяргей Законнікаў	Да	
249	Святлана Басуматрава		
250	Алесь Камароўскі		
251	Раіса Баравікова	Да	Есть
252	Юрка Голуб		
253	Соф’я Шах		
254	Алесь Разанаў	Да	Есть
255	Навум Гальпяровіч		
256	Леанід Якубовіч		
257	Генадзь Пашкоў		
258	Любоў Філімонава		
259	Віктар Ярац		
260	Мікола Пракаповіч		
261	Яўгенія Янішчыц	Да	Есть
262	Пятро Ламан		
263	Алена Руцкая		
264	Зінаіда Дудзюк		
265	Леанід Галубовіч	Да	Есть
266	Галіна Каржанеўская		
267	Міхась Башлакоў		
268	Алесь Каско		
269	Людміла Паўлікава		
270	Алесь Емяльянаў		
271	Алег Мінкін	Да	Есть
272	Васіль Сахарчук		
273	Любоў Тарасюк		
274	Іван Рубін		
275	Уладзімір Марук		
276	Мікола Мятліцкі		
277	Кастусь Жук		

278	Павел Марціновіч	Да	Есть
279	Валянціна Аколава		
280	Алесь Жамойцін		
281	Ірына Багдановіч	Да	Есть
282	Алесь Пісьмянкоў		
283	Леанід Дранько-Майсюк		
284	Сяржук Сокалаў-Воюш		
285	Адам Глобус	Да	
286	Леанід Пранчак		
287	Уладзімір Мазго		
288	Анатоль Сыс	Да	Есть
289	Ала Канапелька		
290	Віктар Шніп		
291	Алесь Аркуш		
292	Галіна Булыка		
293	Алесь Бадак		
	Всего:	79	73

Примечание: Последний столбец иллюстрирует не наличие автора на ресурсе knihi.com вообще, а наличие страниц у авторов, которые были отобраны в первую версию. Другие авторы не проверялись.

Приложение 3. Пример метаданных произведения

Параметр (столбец)	*Источник*	*Пояснение*
authorId	Скрипт	Идентификатор автора
authorName	Скрипт	Имя автора
authorPath	Скрипт	Путь к странице автора на knihi.com
linkId	Скрипт	Идентификатор ссылки на текст
heading	Скрипт	Заголовок родительского блока ссылки. Используется для определения переводов
blText	Скрипт	Текст прямой ссылки
blHref	Скрипт	Прямая ссылка
otherPersonName	Скрипт	Упомянутый автор. Для совместных и переводных изданий
otherPersonHref	Скрипт	Ссылка на упомянутого автора
workName	Скрипт	Название
workHref	Скрипт	Ссылка на страницу текста
Authors	knihi.com	Автор
CreationYear	knihi.com	Год создания
Edition	knihi.com	Издание
FirstPublicationYear	knihi.com	Год первой публикации
LangOrig	knihi.com	Язык оригинала
Pravapis	knihi.com	Орфография
PublicationYear	knihi.com	Год публикации
StyleGenre	knihi.com	Стиль и жанр
Title	knihi.com	Название (заголовок)
Translation	knihi.com	Автор перевода
SectionAuthor	knihi.com	Заголовок родительского блока ссылки
Title2	knihi.com	Подзаголовок
AuthorsTranslated	knihi.com	Автор оригинала
Lang	knihi.com	Язык
Source	knihi.com	Источник
Uploaded	knihi.com	Дата добавления
Year	knihi.com	Год (видимо, создания)
SectionsTheme	knihi.com	Заголовок родительского блока ссылки
Original1	knihi.com	Файл текста
Original1desc	knihi.com	Описание файла
Original1sha1	knihi.com	Хэш файла

FilenameSuffix	knihi.com	Формат
Original2	knihi.com	Дубликат Original1 для дополнительного файла
Original2desc	knihi.com	Дубликат Original1 для дополнительного файла
Original2sha1	knihi.com	Дубликат Original1 для дополнительного файла
MaxChapterLevel	knihi.com	Вложенность разделов для сборников и больших произведений
Notice	knihi.com	Примечание
PreparedBy	knihi.com	Редактор
AuthorsLinked	knihi.com	Упомянутые или связанные авторы
Description	knihi.com	Служебное описание
OriginalCreationYear	knihi.com	Год создания оригинала перевода

Приложение 4. Информация о репозитории проекта

Исходный код, схему разметки и собранные материалы находятся в репозитории на GitHub:

k-nem/bpcorpus // GitHub (https://github.com/k-nem/bpcorpus)