ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР
ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION TECHNOLOGY

DOI 10.51885/1561-4212_2024_1_164
МРНТИ 20.23.21

О.Ю. Кожемякина[1], Н.А. Шашок[2], Э.Д. Кожемякина[3]
Федеральный исследовательский центр информационных и вычислительных технологий, г. Новосибирск, Россия
[1]*E-mail: olgakozhemyakina@mail.ru**
[2]*E-mail: nshashok@alumni.nsu.ru*
[3]*E-mail: kojemyakina.elina2017@yandex.ru*

ОРГАНИЗАЦИЯ ХРАНЕНИЯ КОРПУСОВ ПОЭТИЧЕСКИХ ТЕКСТОВ В ИНФОРМАЦИОННЫХ АНАЛИТИЧЕСКИХ СИСТЕМАХ С УЧЕТОМ СПЕЦИФИКИ ПРЕДМЕТНОЙ ОБЛАСТИ

ПӘНДІК САЛАНЫҢ ЕРЕКШЕЛІГІН ЕСКЕРЕ ОТЫРЫП, АҚПАРАТТЫҚ АНАЛИТИКАЛЫҚ ЖҮЙЕЛЕРДЕ ПОЭТИКАЛЫҚ МӘТІНДЕРДІҢ КОРПУСТАРЫН САҚТАУДЫ ҰЙЫМДАСТЫРУ

ORGANIZATION OF STORAGE OF POETIC TEXTS IN INFORMATION ANALYTICAL SYSTEMS, TAKING INTO ACCOUNT THE SPECIFICS OF THE SUBJECT AREA

Аннотация. *В настоящей работе рассматривается вопрос организации хранения корпусов поэтических текстов в информационных аналитических системах с учетом специфики структуры поэтического текста. В Федеральном исследовательском центре информационных и вычислительных технологий разработана и реализована оригинальная программная система автоматизированного комплексного анализа русских поэтических текстов. Информационная система, как соответствующий компонент программной системы, объединяет разнородную информацию о результатах анализа поэтических текстов. Поэтический текст – структура иерархичная по своей языковой природе, что необходимо учитывать при разработке информационных систем, предназначенных для хранения и обработки текстов на естественном языке. Вопрос иерархии текста равнозначно важен для процесса его анализа и для хранения корпусов текстов. Хранилище текстов является, как правило, центральным компонентом информационных аналитических систем и либо проектируется как база данных, либо представляет собой неструктурированный набор данных. Для экспертов-филологов, работающих с системой, принципиально качество данных, что наиболее достижимо при работе с правильно организованным материалом. В результате концептуального проектирования хранилища корпусов поэтических текстов, с учетом специфики объектов предметной области, обосновано целесообразное использование двух систем хранения и поиска данных: реляционной базы данных для хранения связей между объектами в системе, а также объектов, не являющихся частью корпуса, и хранилища файлов с инструментом полнотекстового поиска в корпусе текстов, что повышает качество анализа текстов и расширяет возможности применения системы в целом.*

Ключевые слова: *обработка текстов на естественном языке, информационные системы, базы данных, хранение файлов.*

Аңдатпа. *Бұл жұмыста поэтикалық мәтін құрылымының ерекшелігін ескере отырып, ақпараттық аналитикалық жүйелерде поэтикалық мәтіндер корпусын сақтауды ұйымдастыру мәселесі қарастырылады. Ақпараттық және есептеу технологияларының Федералды зерттеу*

орталығында орыс поэтикалық мәтіндерін автоматтандырылған кешенді талдаудың өзіндік бағдарламалық жүйесі жасалып, жүзеге асырылды. Ақпараттық жүйе бағдарламалық жүйенің тиісті компоненті ретінде поэтикалық мәтіндерді талдау нәтижелері туралы гетерогенді ақпаратты біріктіреді. Поэтикалық мәтін-лингвистикалық сипаты бойынша иерархиялық құрылым, оны табиғи тілде мәтіндерді сақтауға және өңдеуге арналған ақпараттық жүйелерді әзірлеу кезінде ескеру қажет. Мәтін иерархиясы мәселесі оны талдау процесі үшін және мәтін корпустарын сақтау үшін маңызды. Мәтін қоймасы әдетте ақпараттық аналитикалық жүйелердің орталық құрамдас бөлігі болып табылады және дерекқор ретінде жасалған немесе құрылымдалмаған деректер жиынтығы болып табылады. Жүйемен жұмыс істейтін филолог-сарапшылар үшін дұрыс ұйымдастырылған материалмен жұмыс істеу кезінде қол жеткізуге болатын деректердің сапасы түбегейлі. Поэтикалық мәтіндер корпусының қоймасын тұжырымдамалық жобалау нәтижесінде пәндік аймақ объектілерінің ерекшеліктерін ескере отырып, деректерді сақтаудың және іздеудің екі жүйесін орынды пайдалану негізделеді: жүйедегі объектілер арасындағы байланыстарды, сондай-ақ корпустың бөлігі болып табылмайтын объектілерді және мәтіндер корпусында толық мәтінді іздеу құралы бар файлдарды сақтауды сақтау үшін реляциялық мәліметтер базасы, бұл сапаны жақсартады мәтінді талдау және жалпы жүйені қолдану мүмкіндіктерін кеңейтеді.

Түйін сөздер: *Табиғи тілдегі мәтіндерді өңдеу, Ақпараттық жүйелер, мәліметтер базасы, файлдарды сақтау.*

Abstract. *In this paper, the problem of organizing the storage of poetic text corpuses in information analytical systems is considered, taking into account the specifics of the structure of the poetic text. The original software system for automated complex analysis of Russian poetic texts is developed and implemented in The Federal Research Center for Information and Computational Technologies. The information system, as an appropriate component of the software system, combines heterogeneous information about the results of the analysis of poetic texts. A poetic text is a hierarchical structure by its linguistic nature, what must be taken into account when the information systems for storing and processing texts in natural language is designed. The question of the hierarchy of the text is equally important for the process of its analysis and for the storage of text corpuses. Text storage is, as a rule, the central component of information analytical systems and is either designed as a database or is an unstructured data set. For philologists working with the system, the quality of data is principal, what is most achievable when working with properly organized material. As the result of the conceptual design of the storage of poetic text corpuses, taking into account the specifics of the objects of the subject area, the expedient usage of two data storage and search systems is justified: the relational database for storing links between objects in the system, as well as objects that are not part of the corpus, and the file storage with a full-text search tool in the corpus of texts, what improves the quality analysis of texts and expands the possibilities of using the system as a whole.*

Keywords: *Natural language processing, information systems, databases, file storage.*

Введение. В Федеральном исследовательском центре информационных и вычислительных технологий (далее ФИЦ ИВТ) разработана и реализована оригинальная программная система автоматизированного комплексного анализа русских поэтических текстов [1]. Аналогичных систем, производящих автоматизированный комплексный анализ поэтических текстов на русском языке, в настоящее время не существует, что подтверждено отсутствием публикаций с апробированными данными в соответствующих научных изданиях. Термин «информационная система» в рамках проектирования системы ФИЦ ИВТ определяется как соответствующий компонент программной системы, объединяющий разнородную информацию о результатах анализа поэтических текстов. В [2] подробно представлено концептуальное проектирование, учитывающее задачи и требования для полноценной реализации информационной системы.

Любой текст, даже в виде произвольного неструктурированного сообщения, – структура иерархичная по своей языковой природе, что обязательно надо учитывать при разработке информационных систем, хранящих и обрабатывающих тексты на естественном языке. В поэтических текстах уровни структуры сообщения отображаются в уровни структуры стиха, при этом уровни структуры произвольного сообщения и стиха однозначно сопоставимы. Так, тематика произведения определяется на семантическом и

прагматическом уровнях текста [2, 3]. Вопрос иерархии текста равнозначно важен и для процесса его анализа, и для хранения корпусов текстов. В настоящей работе рассматривается вопрос организации хранения корпусов поэтических текстов в информационных аналитических системах с учетом специфики структуры поэтического текста.

Хранилища и базы данных в системах по сбору, хранению и анализу поэтических текстов. В зарубежных и отечественных разработках есть некоторое количество систем, в которых спроектированы, как компоненты, хранилище или база данных. В ряде исследований авторы сами формируют корпуса текстов, обучающие и тестовые выборки, что, очевидно, имеет смысл для понимания корректности работы алгоритмов. В некоторых работах используются корпуса, собранные другими исследователями, и это, как правило, выверенные и корректные в филологическом аспекте массивы текстов.

Для английского языка, вероятно, существует самое большое количество работ, что связано, безусловно, с распространенностью самого языка.

В работе Дж.М. Фоли (1987) [4] 3182 строки поэмы «Беовульф» (к. VII – н. VIII века) внесены в закодированном виде в базу данных. Применяя статистический анализ данных, автор получает подтверждение ряда гипотез относительно языка поэмы. Также для «Беовульфа» К.Р. Барквистом и Д.Л. Ши (1991) [5] создана база данных с внесенными характеристиками каждого стиха произведения. Основное в этом исследовании – статистический анализ аллитерации, основного звукового приема древнеанглийской поэзии.

Стилистический анализ произведений У. Шекспира проведен в работе Г.С. Донау (1970) [6]: база данных разработана с использованием компьютера, но количество слогов каждого слова стихотворения и паттерны акцентуации вводились вручную. Э. Грин, Т. Бодрумлу и К. Найт (2010) [7] использовали методы машинного обучения с учителем для системы анализа, выполняющей перевод и автоматическое создание стихов; обучающую выборку составили сонеты Шекспира.

К. Барбером и Н. Барбером (1991) [8, 9] собрана вручную база данных из 15942 стихов Дж. Чосера с фонетическими характеристиками, к которым был применен компьютер для получения информации об окончаниях слов и их произношении.

В работе Х. Хирджи (2010) [10] предложена система, производящая фонетическую транскрипцию и слоговое деление. Строки стихотворения транскрибируются, преобразуются в ритмические паттерны, которые сравниваются с заранее определенными образцами в собранном автором корпусе стихов. Таким образом, выявляется вероятность определенной акцентуации, и графический интерфейс отображает отмеченные ударные слоги.

Разработанная коллективом авторов система ZeuScansion (2013) [11] использует словари для определения ударного слога. Процесс начинается с синтаксического анализа, и далее применяется правиловый подход для поиска ударения, при отсутствии слова в словаре используется ближайшее слово. Приоритет для авторов – ритмический паттерн, при этом предполагается разделение слогов. Для тестовой выборки был собран корпус из 759 строк, но проанализирован вручную, и верный результат был показан только в 199 случаях.

В системе SPARSAR, описанной в работе Р. Дельмонте (2013) [12, 13], производится автоматический комплексный анализ поэтических текстов в структуре предложения, строки и строфы, основная задача – изучение стиля. Для работы компонента системы, выполняющего выделение конкретных и абстрактных существительных, используются корпус WordNet (база данных лексики английского языка, содержащая больше 150000 слов) – часть корпуса NLTK (Natural Language Toolkit), платформы для создания программ обработки естественного языка на языке Python.

Web-приложение Metricalizer2 [14] – система, разработанная К. Боббенхаузеном и Б. Хаммерихом и состоящая из нескольких компонентов, производит анализ метрических характеристик немецких стихов. Подсистема анализа корпусов текстов, выделяющая акцунтуацию и рифму и подсчитывющая количество метрических форм, использует корпуса «Freiburger Anthologie» и «Textgrid» [15]. «Freiburger Anthologie» – сборник, содержащий стихи периода 1720-1900 гг., специально отобранные для формирования базы данных. Окончательный вариант корпуса находится в открытом доступе с 1999 г. «Textgrid» – исследовательская группа, занимающаяся поддержкой доступа и обмена информацией в гуманитарных и социальных науках с использованием информационных технологий. Репозиторий Textgrid – это хранилище данных. Таким образом, авторы систем Metricalizer2 и SPARSAR используют готовые текстовые хранилища, что, конечно, не уменьшает ценности их исследований. Авторы Metricalizer2 провели эксперимент с 153 стихами, заявленная точность результата после устранения двух стихотворений по причине проблем, связанных с апострофами, составила 94 %.

Работа П. Герваса (2000) [16] содержит описание системы Gervás Prolog – инструмента, выполняющего анализ стихов на испанском языке. Тестовую выборку составили 64 сонета (все с шестью слогами) шести авторов испанского «золотого века» (приблизительно между 1492 и 1659) объемом 896 стихов. Небольшой корпус данных, но точность анализа заявлена в 99,3 % без учета ошибок, связанных с ситуацией, когда варьируется число слогов у определенного размера, и с современной трактовкой конъюнкции звука *y* в испанской фонетике.

Инструмент, предложенный Б. Наварро-Колорадо (2015, 2016) [17-19], изучает метрики сонетов на испанском языке и производит семантический анализ. Корпус из 5078 сонетов XVI и XVII веков преобразуется в формат TEI [20] для возможности использования в других исследованиях. Тестовая выборка при этом составила 100 сонетов объемом 1400 стихов, показанная точность – 92,3 %; однако точность, полученная при работе экспертов с этой же выборкой, составила 96,2 %.

Названная в честь де Камоэнса программа LuCas, созданная Н. Мамедом и описанная в [21-25], реализованная для стихов на португальском языке, считает слоги и делит текст на стихи и строфы. Тестовая выборка для SAEP составила около 200 строф, но оценивалась не точность, а время отклика системы. В тестовый корпус вошли 12 поэм объемом 197 стихов, точность составила всего 82,2 % с учетом свободного стиха. На уменьшенной тестовой выборке (7 поэм объемом 105 стихов без дисметрических или полиметрических) точность уменьшилась до 77,1 %.

Разработка Д. Роби (1993) [26] – полуавтоматический инструмент для анализа «Божественной комедии» Данте Алигьери на итальянском языке (между 1308 и 1321 гг.). Формат ввода текста не уточняется, разделение слова на слоги не предполагает уточнение ударения. На этом этапе запрашивается вмешательство пользователя, который может вносить сохраняемые в базе данных изменения, отображаемые в дальнейшем для пользователя в числе вариантов.

В исследовании Т.М. Рейнсфорда и О. Скривнера (2014) [27] представлен набор методов, используемых для метрической разметки корпуса стихов, написанных на провансальском (окситанском) языке (диалект Южной Франции), – Lo poema de Boecis («Боэций»). Это анонимный фрагмент, написанный около 1010 года и содержащий 257 стихов, соответствующих метрической схеме из 4 и 6 слогов. Алгоритм перебирает все различные комбинации, пока не находит стих из 10 слогов с акцентом на 4-й и 6-й. Формальная оценка точности анализа не сделана, но найденные некорректные 25 стихов из 257 являются исключениями в тексте.

В. Бодуэн и Ф. Ивон разработали проект Métromètre (1996) [28] для анализа классических французских александрийских стихов (1630-1830 гг.). В систему включены четыре модуля для получения информации о стиле авторов и произведений (2004) [29]: корпус текстов в определенном формате, позволяющий отделять тексты, принадлежащие более чем одному автору; непосредственно сам Métromètre; корпус рифмующихся друг с другом слов; модуль текстовой статистики, назначающий словам лексико-семантический класс. Заявленная точность анализа составила 99,7 %.

Проект Anamètre (2011, 2015) [30-32] включает в себя корпус текстов из более чем 500000 стихов, а также инструменты для сканирования и автоматической маркировки стихов на французском языке без ограничения конкретной формой стихотворения.

К. Сгаллова (1964) [33] собрала корпус из ритмических и метрических данных, извлеченных из 6466 стихов чешского поэта Франтишека Яромира Рубеша, разработала метод их кодирования, а также провела вручную статистический и стилистический анализ. В дополнение был опубликован тезаурус (1999) [34] на основе почти 5000 чешских стихов.

Система KVĚTA создана Р. Ибрагимом и П. Плехачом (2011) [35] для анализа стихотворений, написанных также на чешском языке. KVĚTA стала инструментом для разметки корпуса из более чем 2,5 миллионов стихов (2015) [36, 37]. Точность результата измерялась с помощью полуавтоматически размеченного эталонного корпуса, в котором KVĚTA использовалась для разметки метра, а более сложные случаи размечались экспертом. Вручную авторами размечено 300 корпусов из 25779 стихов, метр определен верно в 99,97 %. При разметке корпуса из 2336435 стихов, точность составила 95,34 %.

Из работ, посвященных текстам на русском языке, несомненно выделяется проект коллектива авторов (А.Е. Поляков, И.А. Пильщиков, М.Б. Бергельсон) «Конкорданс к текстам Ломоносова» [38]. Конкорданс построен на основе корпуса авторских текстов, размеченных структурно, филологически и грамматически. Авторы заявляют о создании открытого интернет-ресурса, включающего [39]: корпус текстов Ломоносова, построенный на основе наиболее авторитетных изданий; биографические, литературно-критические и историко-научные работы о Ломоносове; полный алфавитно-частотный конкорданс к текстам Ломоносова. Непосредственно конкорданс основан на филологически корректной цифровой версии академического полного собрания сочинений и писем Ломоносова в 11-ти томах (1950-1983) [40] и ряде дополнительных изданий [41]. Подготовка корпуса включает: первичную разметку текста для представления в электронной библиотеке; дополнительную структурную разметку и сегментацию текста для корпуса; грамматическую разметку и ее ручную постобработку (снятие омонимии, исправление разборов); преобразование в базу данных; построение конкордансов и других производных. Размеченный конкорданс представляет собой базу данных, из которой можно получать различные виды словарей и проводить объективные исследования авторского языка. Форма базы данных, в свою очередь, дает недоступные в традиционных бумажных словарях возможности: динамический выбор примеров по любым параметрам, динамическую сортировку и группировку, быстрый переход из словаря в корпус текстов, просмотр и выдачу словарной информации в различных форматах, генерацию печатных словарей.

Работа А.В. Козьмина «Автоматический анализ стиха в системе Starling» (2006) [42] посвящена автоматизированному определению метроритмических характеристик русских поэтических текстов и опирается на проект «Автоматизированный лингвостиховедческий анализ русских поэтических текстов», или «Вавилонскую башню» – международный интернет-проект сравнительно-исторического языкознания. Среди компонентов этой информационной системы есть разнообразные словари и базы данных: базы данных этимологий; база данных «Тематическая классификация и распределение фольк-

лорно-мифологических мотивов по ареалам» Ю.Е. Березкина; база данных «Квантитативно-реализационный грамматический словарь современного монгольского языка» С.А. Крылова; словари русского языка, а также несколько программных модулей и морфологический анализатор. Средство управления базами данных – STARLING, с его помощью происходит работа с базами данных: составление и выполнение поисковых запросов, добавление новых записей и организация ссылок. К сожалению, после смерти руководителя проекта С.А. Старостина работы были прекращены.

Проектирование и реализация хранилища корпусов поэтических текстов. Центральным компонентом информационной аналитической системы, как правило, является хранилище текстов. Оно может быть спроектировано как база данных либо являться неструктурированным набором данных. Очевидно, что для экспертов-филологов, работающих с системой, однозначно принципиально качество данных, что, в свою очередь, наиболее достижимо при работе с организованным материалом.

В системе комплексного анализа русских поэтических текстов [1], разработанной в ФИЦ ИВТ, хранилище данных реализовано на данный момент в виде реляционной базы данных и предоставляет возможности поиска текстов по конкретным лексическим единицам, по метаданным, по метроритмическим характеристикам, что позволяет далее проводить комплексный анализ всего поэтического текста. Концептуальное проектирование предметной области для этого хранилища прошло несколько этапов, на каждом из которых схема данных дополнялась или претерпевала структурные изменения для решения возникающих в процессе проектирования задач. При этом выявление основных объектов предметной области, их атрибутов и связей между ними – необходимая информация. Объекты исследуемой предметной области: само произведение, авторы, различные словари, справочники, издательства; каждый из объектов содержит ряд атрибутов.

На первом этапе проектирования модели предметной области системы комплексного анализа основными задачами являлись комплексный лингво-стиховедческий анализ произведений из собранного корпуса текстов, а также осуществление поиска по метаданным и метроритмическим характеристикам поэтического текста, что позволило выделить основные объекты предметной области, представленные в табл. 1.

Таблица 1. Основные объекты предметной области

Объект	Атрибуты
Произведение	1. Название 2. Содержание 3. Авторство 4. Год написания 5. Метроритмические характеристики (строфика, рифмовка и др.) 6. Жанрово-стилевая принадлежность
Автор	1. Фамилия 2. Имя 3. Отчество
Справочник жанров	1. Название жанра
Справочник строфической формы	1. Название строфической формы 2. Схема строфической формы
Справочник рифм	1. Название рифмы
Справочник коллокаций	1. Коллокация
Словарь	1. Лексема

	2. Определение лексемы
Издательство	1. Название издательства 2. Город издания

В рамках предметной области основной связкой между объектами является произведение, которое включает в себя как метаданные (год, авторство, издательство), так и характеристики, используемые для проведения филологического анализа. Связи между объектами показаны на рис 1.

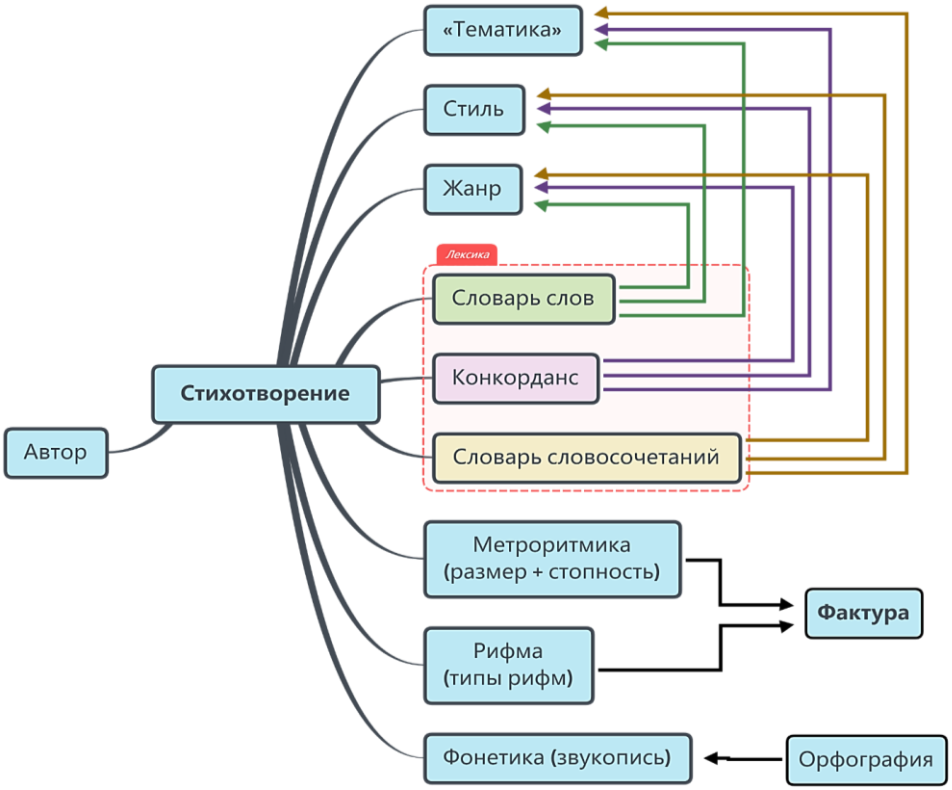

Рисунок 1. Связи между объектами предметной области

В общем случае возможно выделить следующие взаимосвязи объектов рассматриваемой предметной области:

• Произведение содержит в себе атрибуты:

– Метроритмическая принадлежность – рифмовка, строфика;

– Структурная составляющая – состав слов и их частотное употребление;

– Жанрово-стилевая принадлежность – классификация произведения к определенному жанру и стилю.

• Произведение включает в себя метаданные:

– Автор,

– Издательство,

– Год написания (издания).

Очевидно, что кроме общей модели предметной области должна учитываться специфика процесса анализа поэтических текстов, этапы которого определены в наших ранних

работах. Теоретически и практически процесс анализа разделяется на этапы:

- инициализация – формирование корпуса текстов и его предобработка;
- структурный анализ – определение условно низкоуровневых характеристик текста: фонетики и метроритмики;
- семантический анализ – определение смысловых конструкций;
- прагматический анализ – определение жанрово-стилевых особенностей;
- синтез результатов проведенных исследований – определение влияния низших уровней на более высокие, агрегация результатов в удобном для восприятия и поиска виде [2], [43].

Поскольку в информационной системе анализа поэтических текстов ФИЦ ИВТ предполагается одновременное использование более чем одного произведения, объекты предметной области представлены в виде сущностей и связей в соответствии с определенными правилами:

• Один автор может написать несколько произведений. При этом допускается ситуация, когда у одного произведения может быть более одного автора;

• Произведение может выпускаться под несколькими редакциями, которые могут совпадать по году издания;

• Произведение относится только к одному жанру;

• Произведение может иметь ряд слов и словосочетаний, при этом допустимо их повторение в разных произведениях;

• Произведение имеет ряд определяемых метроритмических характеристик (метр, рифма, строфика) [44].

С учетом специфики предметной области и вышеперечисленных правил построена ER-модель, отображающая объекты предметной области, связи этих объектов между собой и ключевые поля (рис. 2).

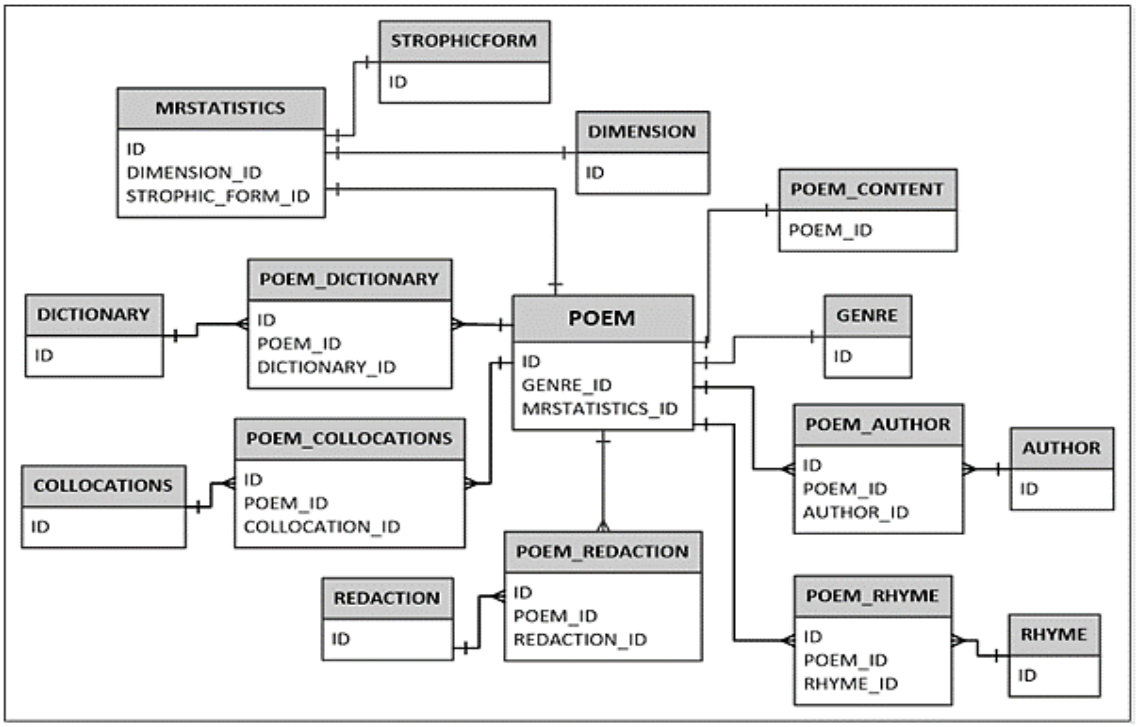

Рисунок 2. ER-модель предметной области

На основании ER-модели была построена схема базы данных, отвечающая поставленным задачам и позволяющая проводить необходимые операции поиска по корпусу.

На этапе проектирования модели предметной области возникла необходимость решать более широкий класс задач, в частности поиск лексем и словоформ с выделением контекста, в котором они находятся внутри произведения. На этапе концептуального проектирования предполагается, что текст поэтического произведения хранится целиком в одной записи в таблице, что не позволяет выводить информацию о контексте по каждой отдельной словоформе или лексеме. Вывод информации о контексте можно реализовать разными способами, например сегментацией текста в базе данных либо хранением номера сегмента без разделения текста на отдельно хранимые фрагменты. Первый вариант позволяет выделить отдельно типы сегментов, которые не являются частью поэтического текста, но являются частью произведения (эпиграфы, сноски). Второй вариант приводит к усложнению использования в поэтическом тексте таких дополнительных сегментов и для поиска, и для организации вывода текста на экран, поэтому для данного этапа была выбрана построчная сегментация с выделением строки текста как объекта предметной области.

Список измененных и добавленных объектов предметной области представлен в табл. 2.

Таблица 2. Измененные и добавленные объекты предметной области

Объект	Тип	Атрибуты
Произведение	Измененный объект	1. Название 2. Авторство 3. Год написания 4. Метроритмические характеристики (строфика, рифмовка и др.) 5. Жанрово-стилевая принадлежность
Строка произведения	Новый объект	1. Текст 2. Тип строки (текст, эпиграф, название части, другое)
Конкорданс	Новый объект	1. Название конкорданса 2. Словоформы в конкордансе и их количество
Словарь словоформ	Новый объект	1. Словоформа 2. Лексема 3. Контексты употребления (фрагменты в произведениях, в которых встречается словоформа, в данном случае строка)

На основании новых объектов предметной области были внесены некоторые изменения в ER-модель данных, представленные на рис. 3.

Проведенная реорганизация модели данных позволяет осуществлять более сложные операции по сравнению с доступными ранее. Однако такой подход привносит некоторые ограничения на изменение и сохранение данных: так, если на первом этапе построения

модели для изменения или удаления отдельного поэтического текста было достаточно одной операции над одним объектом, то на втором этапе изменению или удалению должен подвергаться набор строк, то есть производится несколько операций на нескольких объектах предметной области, что потенциально может привести к нарушению целостности данных. Также такой подход ограничивает сегментирование поэтического текста только строками, несмотря на возможность сегментирования текстов, в том числе, на строфы или предложения.

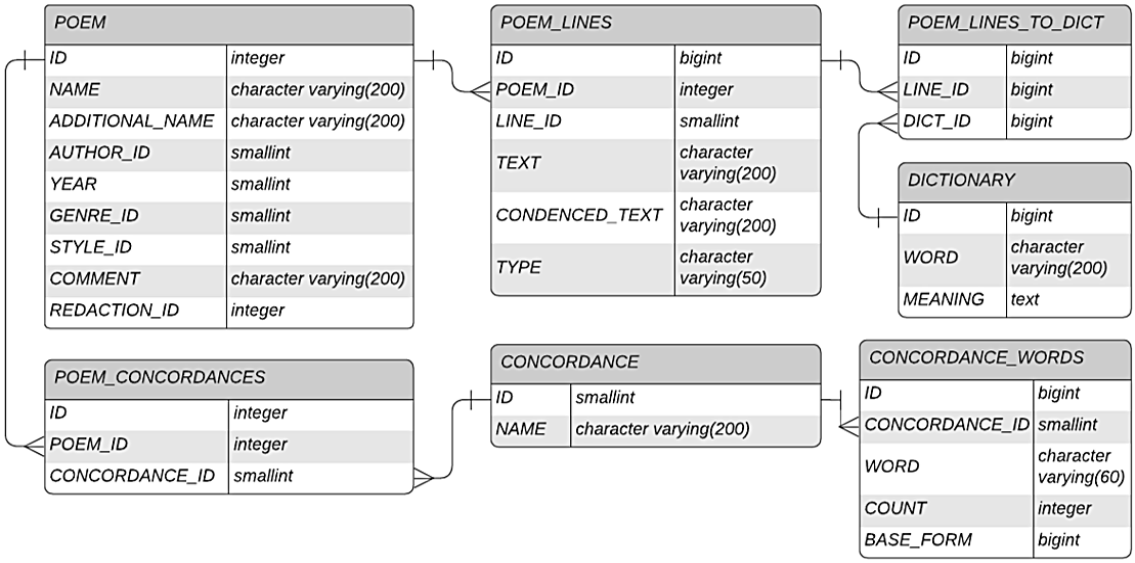

Рисунок 3. Изменения в ER-модели данных

В зависимости от структуры хранения данных устанавливается ограничение на скорость работы, что важно в условиях анализа больших корпусов текстов. Если не производится сегментация, базовые операции с данными производятся достаточно быстро. Однако сегментация позволяет производить более сложные операции, что принципиально для анализа поэтических текстов. Сегментировать текст можно в процессе запроса, но это будет накладывать скоростные ограничения по сравнению с запросами, построенными с учетом изначальной сегментации при хранении. Однако изменение цельного поэтического текста будет намного медленнее при сегментации. По сути, возникает проблема выбора: что важнее для информационной аналитической системы – как можно более быстрый поиск и анализ данных или их изменение.

Принципиальным моментом в проектировании хранилища поэтических текстов является следующее: как именно сегментируется текст, вносимый в систему. Например, в Фундаментальной Электронной Библиотеке [45] сегментация текста производится несколькими способами. Поиск по словам осуществляется в рамках строки или предложения, но поиск в корпусе, где осуществлен вариант поиска в строке, невозможен в рамках предложения. Обратное тоже верно, однако для максимально объективного поиска необходимо предоставить пользователю возможность выбора сегментации: предложение, строка или строфа.

Обозначенная выше проблема выбора сегментации текста возникает одновременно в трех процессах: в хранении, структуризации и поиске данных. Хранение данных в реляционной базе данных априорно предполагает некоторую структуризацию, однако такие данные, как текст, можно хранить в виде файла, внутреннее содержимое которого само

по себе не предполагает никакой структуры, которую, в данном случае, необходимо определить дополнительно. Поскольку выше мы обозначили проблему нарушения целостности данных при подробной сегментации текста в базе данных, рассмотрим также возможность хранения структурированного текста в виде единого файла.

Задача структуризации данных внутри файла, подвергающихся пересекающейся сегментации, рассматривается, в частности, в статье [46] применительно к разметке XML. Подходы, описанные в этой статье, можно применить не только к XML-разметке; представляется возможным их использование также в документах формата JSON либо любого другого формата.

Составленный на примере отрывка из произведения А.С. Пушкина «Евгений Онегин» документ формата JSON с применением пересекающейся сегментации может выглядеть следующим образом:

```
{
    "title": "Письмо Татьяны к Онегину (отрывок из романа «Евгений Онегин»)",
    "poem": {
      "strophes": [
        {
          "lines": [
            {
              "sentences": [
                {
                  "id": 1,
                  "text": "Я к вам пишу — чего же боле?"
                }
              ]
            },
            {
              "sentences": [
                {
                  "id": 2,
                  "text": "Что я могу еще сказать?"
                }
              ]
            },
            …
        ]
      }
    ]
  }
}
```

Если для хранения данных использовать файловое хранилище вместо реляционного, то возникает необходимость в применении инструмента поиска и индексации файлов. Более того, учитывая объем корпуса, по которому производится поиск, следует использовать инструмент, осуществляющий операцию поиска настолько быстро, насколько это возможно. Такими инструментами являются, в частности, Elasticsearch, OpenSearch и Sphinx.

Выводы. Использование инструмента только полнотекстового поиска и индексации в

информационной системе комплексного анализа поэтических текстов на русском языке не является целесообразным из-за наличия других данных текстов кроме содержимого и метаданных – в частности, факт принадлежности текста к нескольким разным словарям или конкордансам не является частью информации собственно о тексте. Из этого следует вывод о рациональности использования двух разных систем хранения и поиска данных: одной для хранения связей между объектами в системе, а также объектов, не являющихся частью корпуса, другой – для поиска в корпусе текстов.

Заключение. Поэтический текст – структура иерархичная на основании природы самого языка, что, очевидно, необходимо учитывать при разработке информационных систем, хранящих и обрабатывающих тексты на естественном языке. Вопрос иерархии текста равнозначно важен для процесса его анализа и для хранения корпусов текстов. Хранилище текстов является, как правило, центральным компонентом информационной аналитической системы и может быть спроектировано как база данных либо представлять собой неструктурированный набор данных. Для экспертов-филологов, работающих с системой, однозначно принципиально качество данных, что, в свою очередь, наиболее достижимо при работе с правильно организованным материалом. С учетом специфики объектов предметной области для концептуального проектирования хранилища корпусов поэтических текстов целесообразно использование двух разных систем хранения и поиска данных: реляционную базу данных для хранения связей между объектами в системе, а также объектов, не являющихся частью корпуса, и хранилище файлов с инструментом полнотекстового поиска в корпусе текстов, что повышает качество анализа текстов и расширяет возможности применения системы в целом.

Список литературы

1. Система комплексного анализа поэтических текстов. – [Электронный ресурс] URL: www.poem.ict.nsc.ru (дата обращения: 26.07.2023).
2. Kozhemyakina O.Yu. Conceptual design of the software system for automated complex analysis of poetic texts // Вычислительные технологии. – 2022. – Т. 27. – № 2. – С. 122-137. – [Электронный ресурс] URL: https://doi.org/10.25743/ICT.2022.27.2.010 (дата обращения: 21.03.2022)
3. Магомедова Д.М. Филологический анализ лирического стихотворения. – М.: Академия, 2004. – 187 с.
4. Foley, J. M. A Computer Analysis of Metrical Patterns in Beowulf. Computers and the Humanities, v. 12, p. 71–80, 1978.
5. Barquist, C. R.; Shie, D. L. Computer Analysis of Alliteration in Beowulf Using Distinctive Feature Theory. Literary and Linguistic Computing, v. 6, n. 4, p. 274–280, 1991
6. Donow, H. S. Prosody and the Computer: A Text Processor for Stylistic Analysis. In: spring joint computer conference, 1970, p. 287–295.
7. Greene, E.; Bodrumlu, T.; Knight, K. Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. In: 2010 conference on empirical methods in natural language processing, 2010, p. 524–533.
8. Barber, C.; Barber, N. The Versification of The Canterbury Tales: A Computer-based Statistical Study. Part I. Leeds Studies in English, v. 21, p. 81–103, 1990.
9. Barber, C.; Barber, N. The Versification of The Canterbury Tales: A Computer-based Statistical Study. Part II. Leeds Studies in English, v. 22, p. 57–83, 1991.
10. Hirjee H. Rhyme, Rhythm, and Rhubarb: Using Probabilistic Methods to Analyze Hip Hop, Poetry, and Misheard Lyrics. – University of Waterloo. – 2010. – [Электронный ресурс] URL: https://uwspace.uwaterloo.ca/bitstream/handle/10012/5419/Hirjee_Hussein.pdf (дата обращения: 21.03.2022).
11. Agirrezabal M., Arrieta B., Astigarraga A., Hulden M. ZeuScansion: a Tool for Scansion of English Poetry // 11th international conference on finite state methods and natural language processing. – 2013. – P. 18–24.
12. Delmonte R. Computing poetry style // Proceedings of 1st International Workshop ESSEM 2013 / CEUR Workshop Proceedngs. – 2013. –№ 1096. – P. 148–155. – [Электронный ресурс] URL:

http://ceur-ws.org/Vol-1096/paper11.pdf (дата обращения: 21.03.2022).
13. SPARSAR. – [Электронный ресурс] URL: https://sparsar.wordpress.com (дата обращения: 21.03.2022).
14. Metricalizer2. – [Электронный ресурс] URL: https://metricalizer.de (дата обращения: 21.03.2022).
15. Freiburger Anthologie, Textgrid. – [Электронный ресурс] URL: https://metricalizer.de/en/about/ (дата обращения: 21.03.2022).
16. Gervás, P. A Logic Programming Application for the Analysis of Spanish Verse // 1st international conference on computational logic. – 2000. – P. 1330–1344.
17. Navarro-Colorado B. A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects // 4th workshop on computational linguistics for literature. – 2015. – P. 105–113.
18. Navarro-Colorado B., Lafoz M. R., Sánchez N. Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation // 9th international conference on language resources and evaluation. – 2016. – [Электронный ресурс] URL: http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf (дата обращения: 21.03.2022).
19. Navarro-Colorado B., Lafoz M. R., Trigueros S. J., Sánchez, N. Compilación y Anotación Métrica de un Corpus de Sonetos del Siglo de Oro // II congreso internacional humanidades digitales hispánicas. – 2015. – [Электронный ресурс] URL: https://hispanismo.cervantes.es/congresos-y-cursos/ii-congreso-internacional-humanidades-digitales-hispanicas-innovacion-0 (дата обращ.: 21.03.2022).
20. Text Encoding Initiative. – [Электронный ресурс] URL: https://tei–c.org/ (дата обращ.: 21.03.2022).
21. Araújo, P. A.; Mamede, N. J. Classi cador de Poemas. In: conferência científica e tecnológica em engenharia, 2002
22. Araújo, P. A. M. Classi cação de Poemas e Sugestão das Palavras Finais dos Versos. 2004. Diss. (Mestrado) – Universidade Técnica de Lisboa
23. Mamede, N.; Trancoso, I.; Araújo, P.; Viana, C. Poetry Assistant. In: 8th international conference on spoken language processing, 2004
24. Mamede, N.; Trancoso, I.; Araújo, P.; Viana, C. An Electronic Assistant for Poetry Writing. In: 9th ibero-american conference on artificial intelligence, 2004, p. 286–294
25. Marques, J. A. D. Sistema de Apoio à Escrita de Poemas. 2008. Diss. (Mestrado) – Universidade Técnica de Lisboa
26. Robey D. Scanning Dante's the Divine Comedy. A Computer–based Approach // Literary and Linguistic Computing. – 1993. – V. 8. – №. 2. – P. 81–84.
27. Rainsford T. M., Scrivner O. Metrical Annotation for a Verse Treebank // 13th international workshop on treebanks and linguistic theories. – 2014. – P. 149–159.
28. Beaudouin V., Yvon F. The Metrometer: a Tool for Analysing French Verse // Literary and Linguistic Computing. – 1996. – V. 11. – № 1. – P. 23–31.
29. Beaudouin V. Mètre en règles // Revue française de linguistique appliquée. – 2004. – V. 9. – P. 119–137.
30. Delente É., Renault R. Annotation automatique de textes versifiés // Schedae. – 2011. – P. 39–52.
31. Delente É., Renault R. Projet Anamètre: Le calcul du mètre des vers complexes // Langages. – 2015. – V. 3. – № 199. – P. 125–148.
32. Delente É., Renault R. Traitement automatique des formes métriques des textes versifiés // 22ème conférence sur le traitement automatique des langues naturelles. – 2015. – P. 432–438.
33. Sgallová, K. Využití moderní techniky při rozboru verše // Česká literatura. – 1964. – V. 12. – № 2. – P. 158–168.
34. Sgallová K. Thesaurus českých meter // Česká literatura. – 1999. – V. 47. – № 3. – P. 286–289.
35. Ibrahim R., Plecháč P. Towards the Automatic Analysis of Czech Verse // Formal Methods in Poetics. – Lüdenscheid: RAM–Verlag, 2011. – P. 295–305.
36. Plecháč, P. Czech Verse Processing System KVĚTA — Phonetic and Metrical Components // Glottotheory. – 2016. – V. 7. – № 7. – P. 159-174.
37. Plecháč P., Kolár R. The Corpus of Czech Verse // Studia Metrica et Poetica. – 2015. – V. 2. – № 1. – P. 107-118.
38. Поляков А.Е., Пильщиков И.А., Бергельсон М.Б. Конкорданс к текстам Ломоносова. – [Электронный ресурс] URL: http://feb–web.ru/feb/lomoconc/abc/ (дата обращения: 21.03.2022).
39. Поляков А.Е., Пильщиков И.А., Бергельсон М.Б. Конкорданс к текстам Ломоносова – концепция и реализация. – [Электронный ресурс] URL: http://www.dialog–21.ru/digests/dialog2009/materials/html/61.htm (дата обращения: 21.03.2022).
40. Ломоносов М. В. Полное собрание сочинений / АН СССР. – М.; Л., 1950-1983.
41. Электронное научное издание «Ломоносов» – [Электронный ресурс] URL: http://feb-

web.ru/feb/lomonos/default.asp?/feb/lomonos/texts/lo0/lo0.html (дата обращения: 21.03.2022).
42. Вавилонская Башня. Проект этимологической базы данных. Русские словари и морфология. – [Электронный ресурс] URL: http://starling.rinet.ru/indexru.htm (дата обращения: 21.03.2022).
43. Барахнин В.Б., Кожемякина О.Ю., Борзилова Ю.С. Проектирование структуры программной системы обработки корпусов поэтических текстов // Распределенные информационно-вычислительные ресурсы. Цифровые двойники и большие данные (DICR-2019): Тр. XVII Междунар. конф. (Новосибирск, 03.12-06.12.2019) / Под ред. О.Л. Жижимова, А.В. Юрченко. – 2019. – Новосибирск: ИВТ СО РАН. – С.102-106. – ISBN: 978-5-905569-14-2. – http://elib.ict.nsc.ru/jspui/bitstream/ICT/4694/20/DICR-2019-V3_p102-106.pdf
44. Барахнин В.Б., Кожемякина О.Ю., Борзилова Ю.С. Оптимизация SQL-запросов на примере работы поискового модуля системы комплексного анализа художественных текстов // Cloud of Science. – 2020. – Т. 7. – № 4. – С. 749-763.
45. ФЭБ "Русская литература и фольклор". – [Электронный ресурс] URL: http://feb–web.ru (дата обращения: 26.07.2023).
46. Schmidt D. The role of markup in the digital humanities // Historical Social Research. – 2012. – V. 27. – № 3. – 125-146

References

1. Sistema kompleksnogo analiza poeticheskih tekstov. – Available at: www.poem.ict.nsc.ru
2. Kozhemyakina O.Yu. Conceptual design of the software system for automated complex analysis of poetic texts // Вычислительные технологии. – 2022. – Т. 27. – № 2. – С. 122-137. – [Электронный ресурс] URL: https://doi.org/10.25743/ICT.2022.27.2.010 (дата обращ.: 21.03.2022)
3. Magomedova D.M. Filologicheskiĭ analiz liricheskogo stihotvoreniya. – M.: Akademiya, 2004. – 187 pp.
4. Foley, J. M. A Computer Analysis of Metrical Patterns in Beowulf. Computers and the Humanities, v. 12, p. 71–80, 1978.
5. Barquist, C. R.; Shie, D. L. Computer Analysis of Alliteration in Beowulf Using Distinctive Feature Theory. Literary and Linguistic Computing, v. 6, n. 4, p. 274–280, 1991
6. Donow, H. S. Prosody and the Computer: A Text Processor for Stylistic Analysis. In: spring joint computer conference, 1970, p. 287–295.
7. Greene, E.; Bodrumlu, T.; Knight, K. Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. In: 2010 conference on empirical methods in natural language processing, 2010, p. 524–533.
8. Barber, C.; Barber, N. The Versification of The Canterbury Tales: A Computer-based Statistical Study. Part I. Leeds Studies in English, v. 21, p. 81–103, 1990.
9. Barber, C.; Barber, N. The Versification of The Canterbury Tales: A Computer-based Statistical Study. Part II. Leeds Studies in English, v. 22, p. 57–83, 1991.
10. Hirjee H. Rhyme, Rhythm, and Rhubarb: Using Probabilistic Methods to Analyze Hip Hop, Poetry, and Misheard Lyrics. – University of Waterloo. – 2010. – Available at: https://uwspace.uwaterloo.ca/bitstream/handle/10012/5419/Hirjee_Hussein.pdf
11. Agirrezabal M., Arrieta B., Astigarraga A., Hulden M. ZeuScansion: a Tool for Scansion of English Poetry // 11th international conference on finite state methods and natural language processing. – 2013. – P. 18–24.
12. Delmonte R. Computing poetry style // Proceedings of 1st International Workshop ESSEM 2013 / CEUR Workshop Proceedngs. – 2013. – № 1096. – P. 148-155. – Available at: http://ceur-ws.org/Vol-1096/paper11.pdf
13. SPARSAR. – Available at: https://sparsar.wordpress.com
14. Metricalizer2. – Available at: URL: https://metricalizer.de
15. Freiburger Anthologie, Textgrid. – Available at: https://metricalizer.de/en/about/
16. Gervás, P. A Logic Programming Application for the Analysis of Spanish Verse // 1st international conference on computational logic. – 2000. – P. 1330–1344.
17. Navarro-Colorado B. A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects // 4th workshop on computational linguistics for literature. – 2015. – P. 105–113.
18. Navarro-Colorado B., Lafoz M. R., Sánchez N. Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation // 9th international conf. on language resources and evaluation. – 2016. – Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf
19. Navarro-Colorado B., Lafoz M. R., Trigueros S. J., Sánchez, N. Compilación y Anotación Métrica de un Corpus de Sonetos del Siglo de Oro // II congreso internacional humanidades digitales hispánicas. –

2015. – Available at: https://hispanismo.cervantes.es/congresos-y-cursos/ii-congreso-internacional-humanidades-digitales-hispanicas-innovacion-0
20. Text Encoding Initiative. – Available at: https://tei–c.org/
21. Araújo, P. A.; Mamede, N. J. Classi cador de Poemas. In: conferência científica e tecnológica em engenharia, 2002
22. Araújo, P. A. M. Classi cação de Poemas e Sugestão das Palavras Finais dos Versos. 2004. Diss. (Mestrado) – Universidade Técnica de Lisboa
23. Mamede, N.; Trancoso, I.; Araújo, P.; Viana, C. Poetry Assistant. In: 8th international conference on spoken language processing, 2004
24. Mamede, N.; Trancoso, I.; Araújo, P.; Viana, C. An Electronic Assistant for Poetry Writing. In: 9th ibero-american conference on artificial intelligence, 2004, p. 286–294
25. Marques, J. A. D. Sistema de Apoio à Escrita de Poemas. 2008. Diss. (Mestrado) – Universidade Técnica de Lisboa
26. Robey D. Scanning Dante's the Divine Comedy. A Computer–based Approach // Literary and Linguistic Computing. – 1993. – V. 8. – №. 2. – P. 81–84.
27. Rainsford T. M., Scrivner O. Metrical Annotation for a Verse Treebank // 13th international workshop on treebanks and linguistic theories. – 2014. – P. 149–159.
28. Beaudouin V., Yvon F. The Metrometer: a Tool for Analysing French Verse // Literary and Linguistic Computing. – 1996. – V. 11. – № 1. – P. 23–31.
29. Beaudouin V. Mètre en règles // Revue française de linguistique appliquée. – 2004. – V. 9. – P. 119–137.
30. Delente É., Renault R. Annotation automatique de textes versifiés // Schedae. – 2011. – P. 39–52.
31. Delente É., Renault R. Projet Anamètre: Le calcul du mètre des vers complexes // Langages. – 2015. – V. 3. – № 199. – P. 125–148.
32. Delente É., Renault R. Traitement automatique des formes métriques des textes versifiés // 22ème conférence sur le traitement automatique des langues naturelles. – 2015. – P. 432–438.
33. Sgallová, K. Využití moderní techniky při rozboru verše // Česká literatura. – 1964. – V. 12. – № 2. – P. 158–168.
34. Sgallová K. Thesaurus českých meter // Česká literatura. – 1999. – V. 47. – № 3. – P. 286–289.
35. Ibrahim R., Plecháč P. Towards the Automatic Analysis of Czech Verse // Formal Methods in Poetics. – Lüdenscheid: RAM–Verlag, 2011. – P. 295–305.
36. Plecháč, P. Czech Verse Processing System KVĚTA — Phonetic and Metrical Components // Glottotheory. – 2016. – V. 7. – № 7. – P. 159-174.
37. Plecháč P., Kolár R. The Corpus of Czech Verse // Studia Metrica et Poetica. – 2015. – V. 2. – № 1. – P. 107-118.
38. Polyakov A.E., Pil'shchikov I.A., Bergel'son M.B. Konkordans k tekstam Lomonosova. – Available at: http://feb–web.ru/feb/lomoconc/abc/
39. Polyakov A.E., Pil'shchikov I.A., Bergel'son M.B. Konkordans k tekstam Lomonosova – koncepciya i realizaciya. – Available at: http://www.dialog–21.ru/digests/dialog2009/materials/html/61.htm Ломоносов М. В. Полное собрание сочинений / АН СССР. – М.; Л., 1950-1983.
40. Lomonosov M. V. Polnoe sobranie sochinenij / AN SSSR. - M.; L., 1950-1983.
41. Elektronnoe nauchnoe izdanie «Lomonosov» – Available at: http://feb-web.ru/feb/lomonos/default.asp?/feb/lomonos/texts/lo0/lo0.html
42. Vavilonskaya Bashnya. Proekt etimologicheskoj bazy dannyh. Russkie slovari i morfologiya. – Available at: http://starling.rinet.ru/indexru.htm
43. Barahnin V.B., Kozhemyakina O.Yu., Borzilova Yu.S. Proektirovanie struktury programmnoj sistemy obrabotki korpusov poeticheskih tekstov // Raspredelennye informacionno-vychislitel'nye resursy. Cifrovye dvojniki i bol'shie dannye (DICR-2019): Tr. XVII Mezhdunar. konf. (Novosibirsk, 03.12-06.12.2019) / Pod red. O.L. Zhizhimova, A.V. Yurchenko. - 2019. - Novosibirsk: IVT SO RAN. – Pp. 102-106. – ISBN: 978-5-905569-14-2. – http://elib.ict.nsc.ru/jspui/bitstream/ICT/4694/20/DICR-2019-V3_p102-106.pdf
44. Barahnin V.B., Kozhemyakina O.Yu., Borzilova Yu.S. Optimizaciya SQL-zaprosov na primere raboty poiskovogo modulya sistemy kompleksnogo analiza hudozhestvennyh tekstov // Cloud of Science. - 2020. – V. 7. – № 4. – Pp. 749-763.
45. FEB "Russkaya literatura i fol'klor". – Available at: http://feb–web.ru
46. Schmidt D. The role of markup in the digital humanities // Historical Social Research. – 2012. – V. 27. – № 3. – 125-146.