Д. В. Степанова (Минск, Беларусь)

РАЗРАБОТКА МЕТОДА ГЕНЕРАЦИИ ЛИНГВИСТИЧЕСКОГО КОРПУСА ИНСТРУМЕНТАМИ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА – ХОД И РЕЗУЛЬТАТЫ МЕЖДУНАРОДНОГО ПРОЕКТА

Описываются этапы выполнения международного проекта временного научного коллектива Минского и Московского государственных лингвистических университетов в рамках 2023-го календарного года. Отдельное внимание уделяется предпосылкам исследования, созданным участниками коллектива несколькими годами ранее. Формулируются предварительные результаты, позволяющие говорить об успешности проекта и обрисовываются перспективы для дальнейшей работы.

Ключевые слова: лингвистический корпус; обработка естественного языка; автоматическая генерация; авторское программное обеспечение; художественная литература; интерпретация текста; русский язык; английский язык; немецкий язык.

The stages of the international project by the research team of Minsk and Moscow State Linguistic Universities in 2023 are described. Special emphasis is placed on the research premises stated by the team members some years earlier. Preliminary results are formulated to indicate the success of the project and prospects for further work are outlined.

Key words: linguistic corpus; natural language processing; automatic generation; special software; fiction; text interpretation; Russian language; English language; German language.

В январе 2023 г. совместным коллективом Минского и Московского государственных лингвистических университетов была начата реализация международного научно-исследовательского проекта «Разработка метода генерации лингвистического корпуса инструментами обработки естественного языка» (№ государственной регистрации 20230455 от 12.04.2023). Научно-исследовательская работа, выполняемая по договору №13/2023 преследует цель определить в рамках корпусного подхода оптимальный с точки зрения достоверности и универсальный метод интерпретации художественных произведений, который позволит извлекать данные для моделирования пространственно-временной и качественной структур произведения, установить черты идиостиля автора.

Задачами исследования являются:

1) определить параметры интерпретации художественных произведений, а именно установить параметры художественной реальности произведения;

2) определить тип создаваемого лингвистического корпуса, необходимого для исследования заданных параметров;

3) отобрать подходящие программные решения, которые определяют не только скорость обработки данных, что немаловажно при объемах современных лингвистических корпусов, но и достоверность получаемых результатов;

4) осуществить предварительную апробацию разрабатываемого метода.

Научно-практический задел для работы был сформирован несколькими годами ранее. В частности, были разработаны алгоритмы анализа пространственно-качественной характеристики художественного произведения на основе частотного анализа и мысленной интерпретации текста [1, с. 45–46]. Далее, данные и сходные алгоритмы были дополнительно успешно апробированы на материале текстов романа Ф. Кафки и его перевода (немецкий и русский языки) [2] и рассказов Дж. Лондона (английский язык) [3].

В 2022 г. также был определен тип лингвистического корпуса, позволяющий проводить исследования произведений художественной литературы: «письменный одноязычный (русский, немецкий, английский языки) литературный художественный исследовательский статический неразмеченный полнотекстовый синхронический корпус» [4, с. 205].

В лаборатории фундаментальных и прикладных проблем виртуального образования Московского государственного лингвистического университета было проведено экспериментальное моделирование базы данных для лингвистического корпуса, которое позволило принять решение в пользу таких программных инструментов, как библиотека обработки естественного языка spaCy и база данных SQL [5]. В качестве «запасного» варианта было обозначено использование базы данных формата XML [6].

Указанный научно-практический задел позволил выйти в 2023 г. на стадию непосредственной программной реализации разработанных экспериментальных моделей.

На сегодняшний момент мы можем констатировать, что полностью выполнены первые три задачи проекта:

1) среди параметров интерпретации художественного произведения отмечены пространственно-качественная характеристика, темпоральная и модальная характеристики, которые активно участвуют в построении художественной реальности произведения. Кроме того, важно обратить внимание на речевой портрет персонажей и способы его реализации в тексте, а также на особенности языка того или иного писателя в принципе для создания формальной модели его идиостиля;

2) лингвистический корпус определен как письменный одноязычный литературный художественный исследовательский статический неразмеченный полнотекстовый синхронический корпус. При оценке качества перевода художественного произведения параметр «одноязычный» может быть заменён на «параллельный»;

3) важным достижением исследовательского коллектива стало создание стабильной бета-версии программного комплекса, состоящего из программы-генератора корпуса и графического интерфейса корпусного менеджера [7, с. 1617–1618], а также функций обработки запросов к базе данных (см. рисунок 1):

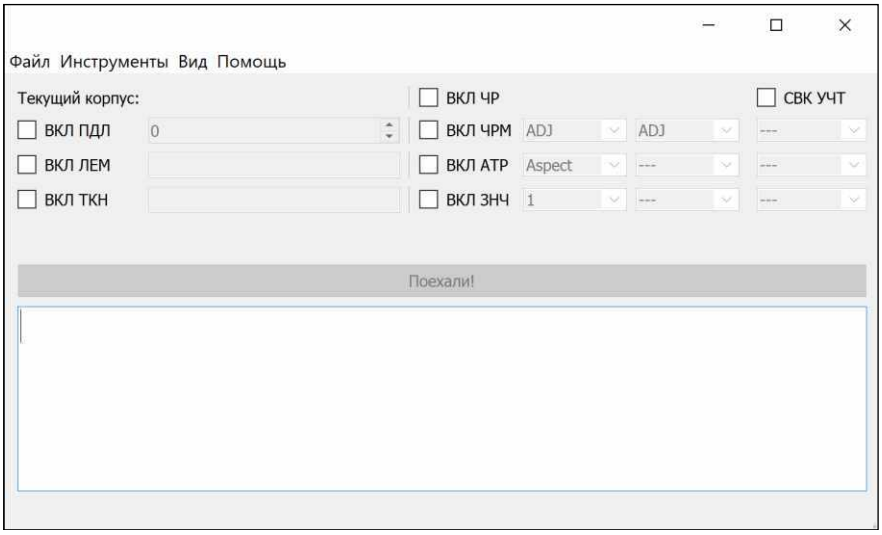

Рисунок 1.
Графический интерфейс пользователя текущей версии корпусного менеджера.

Подчеркнем, что лингвистический корпус создается нами полностью в автоматическом режиме, т. е. без необходимости размечать тексты вручную, тогда как часто «построение полностью размеченного текстового корпуса представляет собой довольно сложный процесс, требующий усилий многих людей» [8 с. 97]. Таким образом, какой-то степени снимается проблема сложности построения лингвистического корпуса.

Приведём ниже примеры результатов вывода базовых запросов для демонстрации некоторых возможностей корпусного менеджера.

Пример 1. Запрос по лемме, т. е. по исходной словоформе парадигмы. Материалом служит сбалансированный корпус текста романа А. К. Дойла “The Lost World” («Затерянный мир»), который насчитывает 4377 предложений и 89873 токена. При этом выводятся все предложения, содержащие заданную лемму, вне зависимости оттого, какая это словоформа. Здесь поиск по лемме “go” (идти) находит словоформы “go”, “went”, “gone” и пр. Приводится номер предложения и его полный текст (см. рисунок 2):

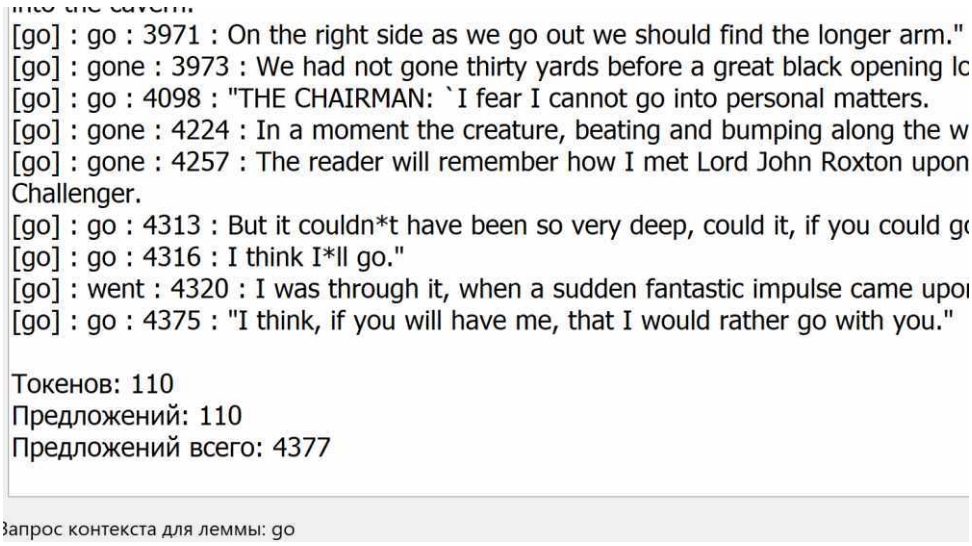

Рисунок 2.
Часть вывода по запросу леммы “go”

Пример 2. Вывод частотного списка определенных частей речи, от самой часто употребительной леммы к самой редкой и с указанием количества ее употреблений в корпусе. Материалом служит уже упомянутый роман А. К. Дойла (см. рисунок 3):

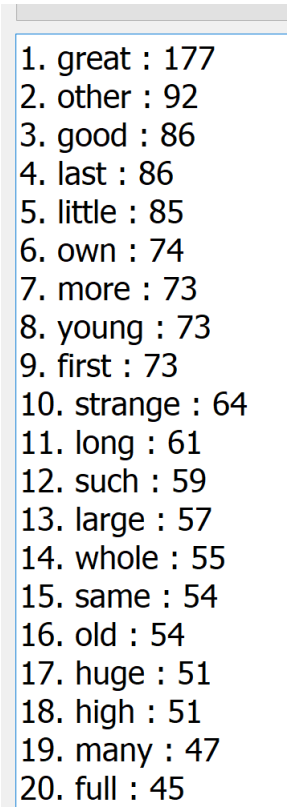

Рисунок 3.
Начало частотного списка прилагательных

Этот список уже позволяет получить самое общее представление о качественных характеристиках художественного произведения.

Для завершения работ по проекту остается провести серию экспериментов, направленных, во-первых, на получение качественных и количественных данных о тексте, преобразованном в базу данных, а во-вторых, на проверку правильности работы программного комплекса.

На наш взгляд, исследование имеет большой потенциал и может быть продолжено дальше – как на материале произведений художественной литературы, так и на материале текстов других жанров.

СПИСОК ЦИТИРУЕМЫХ ИСТОЧНИКОВ

1. Горожанов А. И. Прикладные аспекты анализа и интерпретации текстов (на материале немецкого и русского языков) // А.И. Горожанов, И.А. Гусейнова. Казань Бук, 2021. 208 с. ISBN 978-5-00118-759-2. EDN UNZHVK.
2. Горожанов А. И. Инструментарий автоматизированного анализа перевода художественного произведения / А. И. Горожанов, И. А. Гусейнова, Д. В. Степанова // Вопросы прикладной лингвистики, 2022а. № 45. С. 62-89. DOI 10.25076/vpl.45.03. EDN IWBHQI.

3. Горожанов А. И. Стандартизированная процедура получения статистических параметров текста (на материале цикла рассказов Дж. Лондона "Смок Белью. Смок и Малыш") / А. И. Горожанов, И. А. Гусейнова, Д. В. Степанова // Вестник Минского государственного лингвистического университета. Серия 1: Филология. 2022б. № 4(119). С. 7-13. EDN PXAVUX.
4. Горожанов А. И. Интерпретация художественного произведения: корпусный подход / А. И. Горожанов, Д. В. Степанова // Филологические науки. Вопросы теории и практики. 2022а. Т. 15, № 1. С. 203-208. – DOI 10.30853/phil20220020. EDN TCZLAF.
5. Горожанов А. И. Экспериментальное моделирование базы данных сбалансированного лингвистического корпуса / А. И. Горожанов // Филологические науки. Вопросы теории и практики. 2022. Т. 15. № 10. С. 3382-3386. DOI 10.30853/phil20220563. EDN JHBAVG.
6. Горожанов А. И. Составление сбалансированного корпуса художественного произведения (на материале романов Ф. Кафки) / А. И. Горожанов, Д. В. Степанова // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2022б. № 7(862). С. 31-37. DOI 10.52070/2542-2197_2022_7_862_31. EDN QGIEAQ.
7. Горожанов А. И. Создание лингвистического корпуса на основе инструментов обработки естественного языка: планирование программных решений / А. И. Горожанов // Филологические науки. Вопросы теории и практики. 2023. Т. 16, № 5. С. 1616-1620. DOI 10.30853/phil20230252. EDN BHZCSE.
8. Глазкова А. В. Формирование текстового корпуса для автоматического извлечения биографических фактов из русскоязычного текста / А.В. Глазкова // International Journal of Open Information Technologies. 2019. Т. 7, № 1.С. 97-103. EDN YSKGWL.