УДК 821.161.1

Е.Д. КАЛИНОВСКАЯ
(k.kalinouskaya@yandex.ru)
Московский государственный лингвистический университет

О НЕКОТОРЫХ ВОЗМОЖНОСТЯХ ИНСТРУМЕНТОВ NLP ДЛЯ ПРОВЕДЕНИЯ КОРПУСНЫХ ИССЛЕДОВАНИЙ*

Статья посвящена проблеме автоматического построения лингвистического корпуса. Цель исследования заключается в том, чтобы проанализировать принципы работы программного обеспечения для автоматического построения лингвистических корпусов на примере корпусного менеджера, разработанного в лаборатории фундаментальных и прикладных проблем виртуального образования МГЛУ, а также продемонстрировать его некоторые функциональные возможности на материале оригинального текста романа Артура Конана Дойла «Затерянный мир».

Ключевые слова: *лингвистический корпус, обработка естественного языка, корпусный менеджер, Артур Конан Дойл, роман «Затерянный мир», Python.*

Сбор эмпирических данных является неотъемлемой частью любого современного исследования в области языкознания. До изобретения электронной вычислительной техники ученые-лингвисты вручную собирали и обрабатывали языковой материал. Зачастую этап сбора материала для исследования занимал много лет, а поиск необходимых единиц языка в уже отобранном материале и его обновление являлись весьма трудоёмкими задачами [1].

Развитие компьютерных технологий, появление возможности сканирования и оптического распознавания символов, а также популяризация сети Интернет, привели к бурному росту количества текстов в электронном формате, и дали толчок к возникновению корпусной лингвистики.

По мнению С. Йоханссона, корпусная лингвистика в своём современном виде возникла в США и Западной Европе в 1970-е гг. В это время были открыты первые лаборатории, где лингвисты и программисты стали работать совместно над проблемами автоматической обработки естественного языка [9].

В последующие годы было разработано большое количество корпусов для различных языков: корпус английского языка ‘The Brown corpus’, корпус немецкого языка ‘DWDS’, Корпус испанского языка ‘Corpus de Referencia del Español Actual (CREA)’, Корпус итальянских текстов Болонского университета ‘CORIS’ и т. д. Важнейшую роль в становлении российской корпусной лингвистики сыграло создание Национального корпуса русского языка [6].

Корпусный подход достаточно широко описывается в научной литературе. Так, с точки зрения А.А. Барковича «в парадигме современных исследований языка корпусная лингвистика выделяется методологической универсальностью и эффективностью» [2, с. 5]. Однако А.И. Горожанов и И.А. Гусейнова считают, что уже готовый языковой корпус не является гибким инструментом и не всегда может быть адаптирован для решения узкоспециализированной задачи, как, например, анализ текста литературного произведения с целью получения количественных и качественных данных для заданных параметров [8]. Таким образом, возникает потребность повышения уровня автоматизации построения лингвистических корпусов.

В данный момент на рынке представлено множество библиотек для автоматической обработки естественного языка (NLP), среди них NLTK, spaCy, CoreNLP и мн. др. Библиотеки NLP предлагают

* Работа выполнена под руководством Горожанова А.И., доктора филологических наук, доцента, профессора кафедры грамматики и истории немецкого языка ФГБОУ ВО «МГЛУ».

широкие возможности для корпусных исследований, но в тоже время требуют знание языков программирования. Как правило, это делает невозможным использование данных инструментов лингвистами, не обладающими навыками программирования.

В лаборатории фундаментальных и прикладных проблем виртуального образования Московского государственного лингвистического университета разрабатывается специальное программное обеспечение для корпусных исследований, которое использует библиотеку "spaCy". Это означает, что база данных лингвистического корпуса строится исходя из функциональных возможностей "spaCy" (sentencizer, tokenizer, lemmatizer, morphologizer). Sentencizer определяет границы предложения. Tokenizer разбивает текст на токены (словоформы, знаки препинания и т. д.). Lemmatizer представляет собой компонент конвейера для лемматизации. Morphologizer определяет часть речи токена и прогнозирует его морфологические признаки [10].

Лингвистический корпус хранится в базе данных 'SQLite', которая включает в себя две таблицы – для токенов и для предложений. Таблица предложений состоит из восьми колонок. Колонка 'id' является первичным ключом и создаётся автоматически. Колонка 'sentnum' хранит данные о порядковом номере предложения и связывает две таблицы связью «один ко многим». Колонка 'senttext' содержит текст предложения. Дополнительные пять колонок оставлены пустыми на случай расширения характеристик предложения. Таблица токенов состоит из 12 колонок. Колонка 'id' является первичным ключом и генерируется автоматически. Колонка 'tokennum' предназначена для порядкового номера токена. Колонка 'sent_num' выполняет функцию внешнего ключа и содержит информацию о порядковом номере предложения, в состав которого входит токен. Колонка 'tokentext' предназначена для текста токена, 'tokenpos' – для кода части речи, 'tokenlemma' – для леммы. Колонка 'tokenattr' хранит данные о морфологических характеристиках токена. Как и в первой таблице, дополнительные пять колонок оставлены пустыми на случай расширения характеристик токена [3].

В нашем исследовании в качестве материала для демонстрации функциональных возможностей указанного корпусного менеджера выступает оригинальный текст романа Артура Конана Дойла «Затерянный мир» в формате TXT [7]. Текстовый файл подлежит нормализации – «трансформации в такой вид, который был бы удобен для автоматической обработки с помощью специализированного программного обеспечения» [4, с. 8]. С помощью языка программирования 'Python' была проведена замена серий из двух и более пробелов на одинарный пробел, а также удаление символов переноса строки. Название романа, названия глав и нумерация страниц были удалены вручную.

Загрузка текстового документа и автоматическое заполнение базы данных лингвистического корпуса заняло около двадцати минут.

По классификации, предложенной А.В. Зубовым, полученный корпус является одноязычным (английский язык) литературным художественным исследовательским статическим размеченным полнотекстовым и синхроническим [5].

К базе данных лингвистического корпуса были сделаны запросы на получение следующих параметров:

1) количественное соотношение заданных частей речи в тексте, а именно: существительных, глаголов, прилагательных, местоимений, сочинительных и подчинительных союзов;

2) распределение личных местоимений по грамматическому роду;

3) количество употреблений заданных модальных глаголов, а именно: can (could), may (might), must, should; и список предложений, содержащих данные модальные глаголы.

В результате в тексте романа было найдено 13821 существительное, 9350 глаголов, 6029 прилагательных, 11012 местоимений, 2768 сочинительных союзов и 2457 подчинительных союзов. Полученные данные представлены ниже (см. рис. 1 на с. 83).

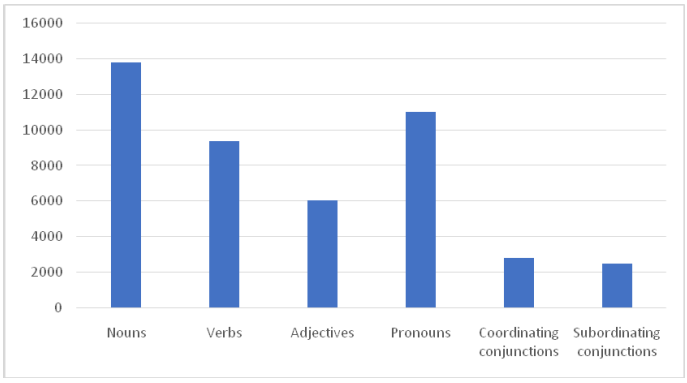

Рис. 1. Количественное соотношение заданных частей речи

Что касается распределения личных местоимений по родам, то в тексте романа было обнаружено 1587 местоимений мужского рода, 84 местоимения женского рода и 1336 местоимений среднего рода (см. рис. 2).

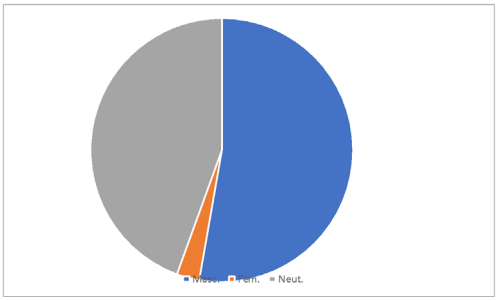

Рис. 2. Распределение личных местоимений по грамматическому роду

Самым частотным модальным глаголом в тексте является глагол ‘can’. В форме настоящего времени он встречается 162 раза и входит в состав 152 предложений, в форме прошедшего времени – 291 раз в 269 предложениях. Модальный глагол ‘may’ в форме настоящего времени используется 130 раз в 120 предложениях, в форме прошедшего времени – 74 раза в 71 предложении. Модальные глаголы ‘should’ и ‘must’ встречаются 141 и 88 раз в 131 и 86 предложениях соответственно (см. рис. 3).

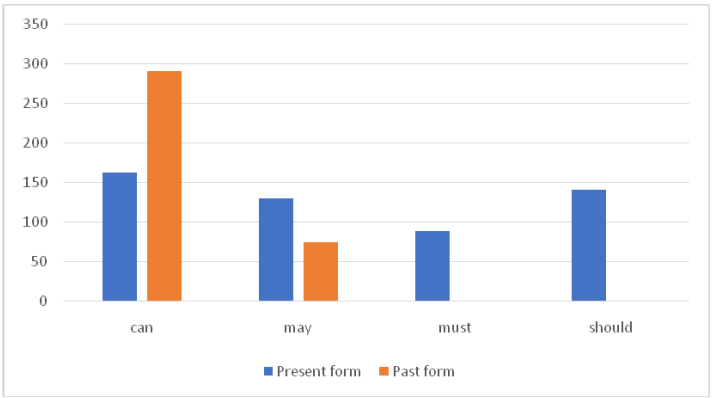

Рис. 3. Количество употреблений заданных модальных глаголов

Подчеркнем, что представленные данные были получены полностью автоматическим путем, без предварительной трудоемкой разметки лингвистического корпуса. Приведённые примеры запросов не исчерпывают возможности разработанного программного обеспечения, которое позволяет получать контексты (предложения) для начальных форм слов, сочетаний частей речи, заданных частей речи с определенными морфологическими признаками и пр. Кроме этого, имеется возможность генерировать частотный список токенов лингвистического корпуса, в том числе и по отдельным частям речи.

Подобное программного обеспечение позволяет решать ряд прикладных задач, связанных с анализом и интерпретацией художественного текста, например: реферирование содержания, определение ключевых локаций повествования, оценка гендерной структуры текста, анализ идиостиля автора и пр.

Перспективным, на наш взгляд, является расширение функционала программы и ее апробация на текстах различных языков.

Литература

1. Баранов А.Н. Корпусная лингвистика // Введение в прикладную лингвистику. М.: Едиториал УРСС, 2021.
2. Баркович А.А. Корпусная лингвистика: специфика современных метаописаний языка // Вестник Томского государственного университета. 2016. № 406. С. 5–13.
3. Горожанов А.И. Экспериментальное моделирование базы данных сбалансированного лингвистического корпуса // Филологические науки. Вопросы теории и практики. 2022. Т. 15. № 10. С. 3382–3386.
4. Горожанов А.И., Гусейнова И.А., Степанова Д.В. Стандартизированная процедура получения статистических параметров текста (на материале цикла рассказов Дж. Лондона «Смок белью. Смок и малыш») // Вестник Минск. гос. лингвистич. ун-та. 2022. № 4(119). С. 7–13.
5. Зубов А.В. Корпусная лингвистика: возможности и перспективы // Русский язык: система и функционирование (к 80-летию профессора П.П. Шубы): мат. III Междунар. науч. конф. (г. Минск, 6–7 апр. 2006 г.): в 2-х ч. Ч. 1. Минск: РИВШ, 2006. С. 22–27.
6. Национальный корпус русского языка: [сайт]. URL: https://ruscorpora.ru/.
7. Doyle A.C. The Lost World: [сайт]. URL: https://www.gutenberg.org/cache/epub/139/pg139.txt.
8. Gorozhanov A.I., Guseynova I.A. Korpusanalyse der Konstituenten Grammatischer Kategorien im Literarischen Text mit Berücksichtigung der Linguoregionalen Komponente // Журнал Сибирского федерального университета. Сер.: Гуманитарные науки. 2020. Т. 13. № 12. С. 2035– 2048.
9. Johansson S. Some aspects of the development of corpus linguistics in the 1970-s and 1980-s // Corpus Linguistics: An International Handbook / ed. by A. Ludeling, M. Kyto. 2008. P. 33–53.
10. spaCy: [сайт]. URL: https://spacy.io/.

EKATERINA KALINOVSKAYA
Moscow State Linguistic University

CONSIDERING THE ISSUE OF THE POTENTIAL OF THE TOOLS OF NLP FOR CORPUS-BASED RESEARCHES

The article deals with the issue of the automatized formation of the linguistic corpus. The aim of the study is to analyze the principles of the work of the software for the automatized formation of the linguistic corpuses at the example of the corpus manager, developed in the laboratory of the fundamental and applied problems of the virtual education in Minsk State Linguistic University, and to demonstrate its functional opportunities based on the original text of the novel "The Lost World" by Arthur Conan Doyle.

Key words: *linguistic corpus, natural language processing, corpus manager, Arthur Conan Doyle, "The Lost World", Python.*