УДК 81'33

Степанова Дарья Валерьевна
кандидат филологических наук, доцент,
доцент кафедры теории
и практики английской речи
Минский государственный
лингвистический университет
г. Минск, Беларусь

Darya Stepanova
PhD in Philology, Associate Professor,
Associate Professor
of the Department of Theory
and Practice of English Speech
Minsk State Linguistic University
Minsk, Belarus
daryastepanova79@gmail.com

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ГЕНЕРАЦИИ ДИНАМИЧЕСКОГО КОРПУСА ТЕКСТОВ СМИ

SOFTWARE PACKAGE FOR GENERATING A DYNAMIC MEDIA TEXTS CORPUS

Статья посвящена проблеме создания динамического лингвистического корпуса точными методами в автоматизированном режиме. Рассматриваются возможность и эффективность использования современных программных инструментов для генерации репрезентативного размеченного корпуса текстов СМИ. Написанные на языке программирования Python коды с применением библиотеки обработки естественного языка spaCy позволили разработать процедуру накопления базы данных корпуса и получить количественные и качественные параметры по заданным запросам.

К л ю ч е в ы е с л о в а: *корпусная лингвистика*; *динамический лингвистический корпус*; *корпусный менеджер*; *база данных*; *обработка естественного языка*.

The article deals with the problem of dynamic linguistic corpus automated creation based on precise methods. The article examines the possibility and efficiency of using modern software tools generating a representative tagged corpus of media texts. The developed Python library programs based on the spaCy natural language processing allow the author of the article to develop the procedure of creating and maintaining a database and to obtain quantitative and qualitative parameters for specified queries.

K e y w o r d s: *corpus linguistics*; *dynamic linguistic corpus*; *corpus manager*; *database*; *natural language processing*.

Одной из актуальных проблем современной прикладной лингвистики является совершенствование методов автоматической обработки естественного языка на основе лингвистических корпусов.

В русле решения данной проблемы в 2023 г. совместным коллективом учреждения образования «Минский государственный лингвистический университет» и федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный

лингвистический университет» был реализован международный научно-исследовательский проект «Разработка метода генерации лингвистического корпуса инструментами обработки естественного языка» (№ государственной регистрации 20230455 от 12.04.2023).

В рамках данного проекта проведено исследование, целью которого является создание в автоматизированном режиме и апробация репрезентативного размеченного динамического корпуса текстов для решения различных лингвистических задач.

Для достижения поставленной цели необходимо решить следующие задачи:

1) определить способ разметки большого массива текста с помощью библиотеки обработки естественного языка spaCy высокоуровневого языка программирования Python, которая в обычном случае допускает загрузку текстового файла объемом не более 1 МБ;

2) построить в автоматизированном режиме корпус текстов объемом не менее 1 млн токенов;

3) протестировать созданный корпус на предмет целостности базы данных;

4) посредством запросов извлечь из полученного корпуса данные о тематическом наполнении публикаций.

Материалом для проведения данного исследования послужили тексты статей немецкоязычного периодического издания Spiegel, который в силу широкого тематического охвата и высокого качества языка публикаций является популярным объектом различных лингвистических исследований [1; 2; 3; 4].

Методами исследования выступают алгоритмизация (в части построения программного решения), автоматический программный анализ и синтез полученных в результате данных.

Для решения первой задачи нами были внесены модификации в спроектированный ранее модуль автоматической генерации лингвистического корпуса по правилам библиотеки обработки естественного языка spaCy [5, с. 1617]. Решение состоит в том, чтобы последовательно создавать корпус из N текстовых массивов (файлов txt) объемом менее 1 МБ, т. е. пополнять корпус новым лингвистическим материалом.

Решение второй задачи потребовало организовать процедуру регулярного отбора текстов рассматриваемого периодического издания. Данная процедура осуществлялась в несколько этапов: определение рубрики-источника, автоматизация отбора материала, разработка технологии отбора заголовков статей, построение базы данных лингвистического корпуса.

В качестве рубрики-источника был выбран раздел Schlagzeilen (заголовки), который ежедневно пополняется десятками новостных статей различных тематик, например, обзор событий в глобальной сети Интернет, экономика, политика, культура, спорт, международные события, наука, история и другие.

В целях исключения «ручного» режима копирования текстов была написана программа на языке программирования Python, которая в автоматизированном режиме осуществляет отбор заголовков статей по заданным параметрам и создает нормализованный текстовый массив [6, с. 8–9]. При этом следует обращать внимание на то, чтобы размер файла не превышал 1 МБ.

Для обеспечения возможности построить динамический корпус и регулярно пополнять лингвистическую базу новыми текстами была разработана технология, которая при первом запуске программы предусматривает отбор в автоматизированном режиме заголовков журнала, а при втором и последующих запусках – добавляет только новые заголовки для формирования текстового массива, состоящего из новых текстов. Таким образом, после каждого запуска программы вносятся дополнения в общую базу с заголовками и формируется корпус текстов статей по новым заголовкам. Отбор текстов осуществляется без метаданных и без разделения на абзацы.

Полученные текстовые файлы один за другим подвергаются автоматической обработке генератором корпуса. Количество файлов, подлежащих процедуре обработки за один цикл, не ограничено. Получаемый в процессе такого формирования данных лингвистический корпус может иметь значительный объем, превышающий один миллион токенов. База данных корпуса состоит из двух таблиц: таблицы токенов и таблицы предложений [7, с. 3383–3384].

Для решения следующей задачи – тестирования лингвистического корпуса на предмет целостности – была проанализирована база данных в части присвоения числовых идентификаторов. В случае, если последнее предложение корпуса имеет номер N, то при добавлении предложений из нового текстового массива очередному предложению должен быть присвоен номер N+1. То же должно быть верно и для токенов, но с той разницей, что при соотнесении токена с предложением в таблице предложений все номера должны быть корректно синхронизированы.

Приведем в качестве пояснения фрагмент кода для добавления текста в соответствующий лингвистический корпус (листинг):

```
    # Заполнение новой db из нескольких файлов
    elif " " not in self.ui.lineEdit_corpusName.text() and self.weCanStartMult
== True:
        dbName = self.ui.lineEdit_corpusName.text() + ".db"
        self.createANewDB(dbName)
        ## Структура таблицы
        table01 ="""
                CREATE TABLE IF NOT EXISTS sents (
                id integer PRIMARY KEY,
                sentnum integer NOT NULL,
                senttext text NOT NULL,
                sentoption01 text DEFAULT 'NONE',
                sentoption02 text DEFAULT 'NONE',
                sentoption03 text DEFAULT 'NONE',
                sentoption04 text DEFAULT 'NONE',
                sentoption05 text DEFAULT 'NONE'
            );
            """
        ## Структура таблицы
        table02 = """
            CREATE TABLE IF NOT EXISTS tokens (
            id integer PRIMARY KEY,
            tokennum integer NOT NULL,
            sent_num integer NOT NULL,
            tokentext text NOT NULL,
            tokenpos text,
            tokenlemma text,
            tokenattr text,
            tokenoption01 text DEFAULT 'NONE',
            tokenoption02 text DEFAULT 'NONE',
            tokenoption03 text DEFAULT 'NONE',
            tokenoption04 text DEFAULT 'NONE',
            tokenoption05 text DEFAULT 'NONE',
            FOREIGN KEY (sent_num) REFERENCES sents (id)
        );
        """
        self.createATable(table01, dbName)
        self.createATable(table02, dbName)
        # Открывает библиотеку для всех файлов один раз
        nlp = spacy.load(self.ui.comboBox_selectLang.currentText())
        # Счетчик предложений устанавливается на 1 или на текущую позицию,
если идет добавление в имеющуюся БД
        conn = sqlite3.connect(dbName)
        indexSent = int(self.getStrSQL(conn.execute("SELECT COUNT (*) FROM
sents")))
        if indexSent == 0:
            counterSent = 1
        else:
            counterSent = indexSent + 1
        conn.close()
        # Открывает папку с текстовыми файлами для корпуса
        fileNames = os.listdir(self.TXTNameFolder)
        # Прибавочный индекс для счета токенов
        index = 0
        # Перебирает все файлы из папки
```

```
        for file in fileNames:
            fileTemp = open(self.TXTNameFolder + "/" + file, "r",
encoding="utf-8")
            text = fileTemp.read()
            fileTemp.close()
            # Все одинарные кавычки заменяются на звездочки
            text = text.replace("'", "*")
            # Заполнение таблиц базы данных из текстов по очереди
            doc = nlp(text)
            # Блок полосы прогресса - пока для каждого файла отдельно,
показано имя текущего файла
            self.cancelled = False
            self.progress = QtWidgets.QProgressDialog(file + " сборка...",
"Стоп", 0, len(doc))#, self.ui.action_freq)
            self.progress.setWindowModality(QtCore.Qt.WindowModal)
            self.progress.setMinimumDuration(10)
            i = 0
            # Подключение к базе данных
            conn = sqlite3.connect(dbName)
            # Запрос текущего количества токенов в базе
            index = int(self.getStrSQL(conn.execute("SELECT COUNT (*) FROM
tokens")))
            # Цикл для перебора каждого токена
            for token in doc:
                # Блок полосы прогресса
                self.progress.setValue(i)
                i += 1
                if self.progress.wasCanceled():
                    self.cancelled = True
                    return
                # Все кавычки в лемме заменяются на описательные слова
                if token.text == "'":
                    lemma = "quote"
                elif token.text == '"':
                    lemma = "doublequote"
                else:
                    lemma = token.lemma_
                if token.is_sent_start:
                    conn.execute("INSERT INTO sents VALUES(NULL, %d, '%s',
NULL, NULL, NULL, NULL, NULL)" % (counterSent, token.sent))
                    conn.commit()
                    conn.execute("INSERT INTO tokens VALUES(NULL, %d, %d,
'%s', '%s', '%s', '%s', NULL, NULL, NULL, NULL, NULL)" % (token.i+index,
counterSent, token.text, token.pos_, lemma, token.morph))
                    conn.commit()
                    if token.is_sent_end:
                        counterSent += 1
                elif token.is_sent_end:
                    conn.execute("INSERT INTO tokens VALUES(NULL, %d, %d,
'%s', '%s', '%s', '%s', NULL, NULL, NULL, NULL, NULL)" % (token.i+index,
counterSent, token.text, token.pos_, lemma, token.morph))
                    conn.commit()
                    counterSent += 1
                else:
                    conn.execute("INSERT INTO tokens VALUES(NULL, %d, %d,
'%s', '%s', '%s', '%s', NULL, NULL, NULL, NULL, NULL)" % (token.i+index,
counterSent, token.text, token.pos_, lemma, token.morph))
        conn.commit()
```

Генерация корпуса проводилась в период с 31.05.2023 г. по 22.06.2023 года. Общий объем текстового массива составил 6,58 МБ в 21 файле размерами от 82 КБ до 902 КБ. Количество включенных статей журнала Spiegel оказалось равным 3015.

После автоматического преобразования текстового массива в базу данных был получен лингвистический корпус объемом 63 578 предложений и 1 107 262 токена. Техническая апробация базы данных лингвистического корпуса показала, что порядковые номера предложений и токенов не имеют дублирований или разрывов. Таким образом, задача по разработке корпуса текстов объемом не менее 1 млн токенов в автоматизированном режиме была успешно выполнена.

В рамках решения задачи по извлечению данных о тематическом наполнении публикаций посредством запросов из полученного корпуса текстов был построен частотный словарь имен собственных, который показал следующие результаты (приводятся первые 25 позиций словаря с указанием количества их употреблений в созданном корпусе):

1. Ukraine : 1173
2. Deutschland : 1128
3. Russland : 974
4. China : 550
5. USA : 538
6. Trump : 523
7. SPIEGEL : 486
8. Berlin : 360
9. Scholz : 311
10. SPD : 310
11. Moskau : 308
12. Europa : 292
13. Putin : 282
14. Kiew : 263
15. Bayer : 252
16. Biden : 247
17. Selenskyj : 235
18. Twitter : 218
19. München : 217
20. EU : 190
21. FDP : 180
22. Erdoğan : 177
23. Frankreich : 171
24. CDU : 169
25. Türkei : 168

Полученные результаты позволяют сделать вывод о наиболее актуальных вопросах общественной и политической жизни, освещаемых в рассматриваемом периодическом издании.

В качестве дополнительных возможностей созданного корпуса текстов рассмотрим некоторые результаты анализа данных, полученных посредством формирования различных запросов. К примеру, в целях определения доли токенов, которым присуще свойство «sub», т. е. «употребление в конъюнктиве», построим запрос к базе данных:

SELECT COUNT (*) FROM tokens WHERE tokenattr LIKE '%=sub%'

По результату данного запроса было получено 12 627 лексических единиц, что составляет 1,14% от их общего количества в корпусе. Таким образом, суммарная доля глаголов, которые употребляются в наклонениях конъюнктив I и конъюнктив II, находится на уровне одного процента от всех токенов. При этом общее количество глаголов в корпусе составляет 87 739 единиц.

В рамках решения задачи по установлению особенностей употребления лексических единиц с атрибутом «женский род» построим запрос к базе данных:

SELECT COUNT (*) FROM tokens WHERE tokenattr LIKE '%=fem%'

Аналогичным образом сформируем запросы для установления единиц с атрибутами «мужской род» и «средний род». В результате были получены следующие данные: доля лексических единиц с атрибутом «женский род»: 178 690 (16,14 %); для мужского: 179 435 (16,21 %); для среднего: 118 799 (10,73 %). Полученные данные позволяют сделать вывод о практически полном равенстве «мужского» и «женского» в текстах статьей рассматриваемого издания Spiegel.

Приведенные данные свидетельствуют об эффективности предлагаемого способа отбора текстов для создания лингвистического динамического корпуса и его использования для решения различных задач.

Таким образом, в результате проведенного исследования был построен и апробирован динамический лингвистический корпус новостных текстов периодического журнала Spiegel объемом 1 107 262 токена. При этом удалось решить проблему преобразования текстового файла объемом более 1 МБ средствами библиотеки spaCy. Полученная база данных, которая составляет основу лингвистического корпуса, не обнаружила ошибок целостности. Для дополнительной апробации разработки был сгенерирован частотный словарь имен собственных, который позволяет получить ясное представление о тематическом наполнении публикаций. Также были продемонстрированы возможности корпуса для анализа различных язы-

ковых явлений. В качестве перспективного направления исследования представляется целесообразным разработать схожие корпуса текстов на других языках для проведения сопоставительных исследований.

ЛИТЕРАТУРА

1. *Соловьева, В. Э.* Постколониальная тематика на страницах европейской прессы / В. Э. Соловьева // Журналистика, мультимедиа: информационный и социокультурный потенциал : материалы V Всероссийской науч.-практ. конф., посвящ. памяти Г. М. Соловьева, Краснодар, 02 дек. 2022 г. – Краснодар : Кубан. гос. ун-т, 2023. – С. 310–315.
2. *Езан, И. Е.* Лингводискурсивный корпусный анализ лексики пандемии коронавируса в онлайн-версии журнала Der Spiegel / И. Е. Езан, Е. А. Ковтунова, Л. Н. Григорьева // Вестн. Санкт-Петербург. ун-та. Язык и литература. – 2022. – Т. 19, № 4. – С. 760–779.
3. *Мишин, А. В.* Языковой образ России в жанре интервью (на материале рубрики «Spiegel-Gespräch» журнала «Der Spiegel») / А. В. Мишин // Вестн. Удмурт. ун-та. Сер.: История и филология. – 2022. – Т. 32, № 4. – С. 822–827.
4. *Василькова, М. В.* Стилистика проблемного интервью (на материале журнала «Spiegel») / М. В. Василькова // Наука – образованию, производству, экономике : материалы 72-й Региональной науч.-практ. конф. преподавателей, науч. сотрудников и аспирантов, Витебск, 20 фев. 2020 г. / редкол.: И. М. Прищепа (гл. ред.) [и др.]. – Витебск : Витеб. гос. ун-т им. П. М. Машерова, 2020. – С. 110–112.
5. *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.
6. *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 // Вестн. МГЛУ. Сер. 1., Филология. – 2022. – № 4 (119). – С. 7–13.
7. *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.

Поступила в редакцию 28.11.2023