ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2023. Том 16. Выпуск 5. С. 1616-1620 | 2023. Volume 16. Issue 5. P. 1616-1620
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Создание лингвистического корпуса на основе инструментов обработки естественного языка: планирование программных решений

Горожанов А. И.

Аннотация. Целью исследования является построение модели лингвистического корпуса, генерация которого происходит по правилам библиотеки обработки естественного языка spaCy. Научная новизна заключается в том, что в рамках гуманитарного исследования применяется метод моделирования, который сочетается с корпусным подходом и учитывает технологический (программный) компонент уже на стадии целеполагания. В ходе работы, во-первых, была определена общая структурная модель лингвистического корпуса в виде последовательности блоков и сформулированы типовые запросы к его базе данных, во-вторых, построена модель интерфейса корпусного менеджера, способного реализовать эти типовые запросы, и, в-третьих, проведен анализ предложенной модели с помощью отдельных мини-программ, позволяющих оценить степень технической реализуемости запросов и их практическую ценность. На этой стадии в качестве языкового материала были привлечены текстовые массивы художественных произведений немецкоязычных (Ф. Кафка, Э. М. Ремарк) и англоязычных (А. К. Дойл, Дж. Оруэлл) писателей. Полученные результаты показали, что построенная модель имеет ряд достоинств при ограниченном количестве недостатков, обладает параметром гибкости в плане дальнейшего развития и может быть программно реализована в краткосрочной перспективе.

EN

Building a linguistic corpus based on natural language processing tools: Planning software solutions

Gorozhanov A. I.

Abstract. The paper is aimed at building a model of a linguistic corpus, which is generated according to the rules of the spaCy natural language processing library. Scientific novelty lies in the fact that within the framework of humanities research, the method of modelling is used, which is combined with a corpus approach and takes into account the technological (software) component at the very stage of goal setting. In the research, firstly, a general structural model of a linguistic corpus as a sequence of blocks was determined and standard queries to the database were formulated; secondly, a model of the corpus manager interface able to implement these standard queries was built; thirdly, an analysis of the proposed model with the help of mini-programs that allow assessing the degree of technical feasibility of the queries and their practical value was conducted. At this stage, text arrays of fictional works by German-speaking (F. Kafka, E. M. Remarque) and English-speaking (A. C. Doyle, G. Orwell) writers were involved as linguistic material. The obtained results showed that the constructed model has a number of advantages with a limited number of disadvantages, is flexible in terms of further development and can be programmatically implemented in the short term.

Введение

Графический интерфейс пользователя необходим каждому приложению для того, чтобы оно стало рабочим инструментом для специалиста, не обладающего навыками программирования.

Проблема разработки такого интерфейса является актуальной не только с чисто технической точки зрения, но и с позиции эргономики получаемого продукта, о чем свидетельствует значительный интерес к этой теме, проявляемый в современной предметно-специальной литературе (Бойко, Легалов, Зыков, 2022; Читалов, 2022; Бакаев, Разумникова, 2017).

Лингвистический корпус как программное приложение тем более должен иметь удобную систему управления с помощью привычных всем виджетов, так как гуманитарные исследователи редко обладают навыками написания кода и не могут самостоятельно напрямую работать с современными программными инструментами, какими, например, являются библиотеки обработки естественного языка.

Научная статья (original research article) | https://doi.org/10.30853/phil20230252

Необходимость в моделировании интерфейсов управления лингвистическим корпусом как «шаблонного» процесса, учитывающего не только технические, но также и лингвистические параметры, обуславливает актуальность нашего исследования, в ходе которого предполагается решить следующие задачи:

1. Построить общую структурную модель лингвистического корпуса и описать характер запросов к созданной ранее реляционной базе данных как части лингвистического корпуса.
2. Построить модель графического интерфейса пользователя корпусного менеджера как системы управления базой данных.
3. Провести оценку модели лингвистического корпуса.

Практическая значимость работы состоит в том, что ее результаты позволяют перейти непосредственно к написанию программного кода корпусного менеджера без необходимости его значительной корректировки, что существенно сокращает время от начала этапа планирования до завершения этапа внедрения лингвистического корпуса в сферу реальных исследований.

Методами исследования послужили моделирование в части разработки общей модели лингвистического корпуса и модели графического интерфейса пользователя корпусного менеджера и анализ в части оценки модели лингвистического корпуса в целом.

Лингвистическим материалом послужили аутентичные тексты произведений Ф. Кафки («Замок»), Э. М. Ремарка («На Западном фронте без перемен»), А. К. Дойла («Затерянный мир») и Дж. Оруэлла («1984»).

Теоретическую базу исследования составили труды отечественных и зарубежных ученых, посвященные корпусным исследованиям (Tsujii, 2021; O'Neill, Welsh, Smith et al., 2021; Fonseca, Guelpeli, De Souza Netto, 2021; Malyuga, McCarthy, 2021), в том числе и в плане разработки технических решений (Писарик, 2021; Горожанов, 2022; Горожанов, Степанова, 2022). Предполагается, что настоящая работа послужит вкладом в формирование методологической базы лингвистических исследований, которые опираются на корпусный подход и занимаются вопросами построения корпусов с помощью современных программных инструментов.

Обсуждение и результаты

В начале нашей работы необходимо было определить компонентный состав модели или, другими словами, построить общую структурную модель, которая бы позволила наглядно представить процесс функционирования лингвистического корпуса.

Двумя первыми компонентами здесь являются библиотека обработки естественного языка spaCy и работающий на ее основе генератор корпуса, который предполагает наличие минимального графического интерфейса пользователя. Генератор получает файл TXT с исходным текстовым массивом и формирует реляционную базу данных корпуса, которая состоит из двух таблиц: таблицы предложений и таблицы токенов (Горожанов, 2022, с. 3383-3384).

Оперирование базой данных предполагается осуществлять с помощью специальной системы управления базой данных (СУБД) – корпусного менеджера, который должен иметь сложный графический интерфейс пользователя для построения большого количества запросов к базе данных. Собственно, база данных и корпусный менеджер и образуют собой лингвистический корпус в «классическом» понимании, который функционирует уже без обращения к библиотеке обработки естественного языка.

Представим сказанное в виде схемы (см. Рис. 1).

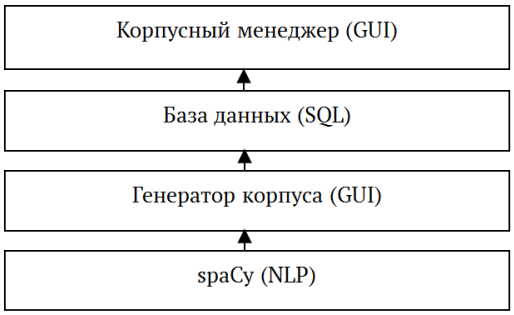

Рисунок 1. *Общая структурная модель*

Здесь GUI – графический интерфейс пользователя; SQL – специальный язык для оперирования реляционной базой данных; NLP – обработка естественного языка.

Рассмотрим возможные запросы. Поскольку база данных состоит из двух таблиц, выделим отдельные запросы к таблице предложений и к таблице токенов. Предложения имеют только одну характеристику – порядковый номер. Поэтому прямой запрос к этой таблице будет иметь целью вывод предложения или последовательности предложений по их номеру. Общее количество предложений в корпусе должно быть указано в интерфейсе корпусного менеджера. К запросам, обращенным исключительно к таблице токенов, можно отнести построение частотного списка токенов – как общего, так и по отдельным частям речи.

Наиболее распространенными запросами с технической точки зрения будут смешанные запросы, в рамках которых корпусный менеджер должен будет обращаться сразу к обеим таблицам. Предполагается, что результатом таких запросов станут контексты (предложения), содержащие токены с заданными параметрами.

Первым из таких запросов будет поиск по токену. При этом менеджер выведет пронумерованные предложения, в которых встретится токен в той форме, в какой он был введен пользователем. В таблице токенов будет происходить поиск соответствия, а затем вывод предложения, соотнесенного с этим токеном. Если заданный токен встречается в одном и том же предложении n > 1 раз, то это предложение будет выведено n раз. В конце списка предложений должна помещаться статистическая справка с информацией о количестве найденных токенов и предложений *без дублирования*.

Сходным образом должен осуществляться поиск по лемме (здесь: исходной форме токена), причем следует предусмотреть ввод нескольких лемм для поиска контекстов, отражающих заданную тематическую группу, например «мир, дружба, торговля, помощь» для русских текстов.

Далее необходимо будет предусмотреть поиск определенных частей речи. Например, только числительных или прилагательных. Результатом вывода также должны быть предложения. Целесообразным для статистического анализа представляется запрос по поиску набора до трех частей речи, например имен собственных (по правилам spaCy они выделяются в отдельный класс) и числительных и т. д.

В таком варианте пользователь получает данные о количестве предложений, в которых встречаются одновременно все заданные части речи. Следующими двумя вариантами запросов будут поиск по атрибутам (морфологическим признакам) частей речи и поиск по значениям этих атрибутов.

Здесь также уместно предусмотреть до трех значений параметра. Например, пользователь сможет вывести предложения, в которых есть токены с атрибутом «лицо», «число», «наклонение», «падеж» и пр. В случае поиска по значениям это может быть женский, мужской или средний род, именительный падеж, сравнительная степень и т. д.

Несмотря на то, что практическая ценность указанных запросов может не казаться в данный момент очевидной, они предоставляют лингвисту-исследователю достаточно гибкий набор сочетаний, который может быть полезным при анализе текстов различных жанров. Например, при помощи значения «женский/мужской род» можно определить гендерную окраску текста, а значение «настоящее/прошедшее» поможет выявить его темпоральную доминанту.

Следующую группу запросов мы определим как комплексные запросы или запросы по сочетанию критериев.

В первую очередь назовем здесь поиск по лемме либо по токену с дополнительным указанием части речи, например англ. square с включенным признаком «часть речи = существительное». Этим мы отделим существительные square (‘квадрат’) от идентичных им по написанию прилагательных (‘квадратный’). Далее полезной кажется комбинация леммы/токена с атрибутами и значениями (до трех параметров). Например, для немецкого языка возможным будет поиск заданного глагола в определенном лице, числе и времени (наклонении, залоге и пр.) или поиск заданного прилагательного в сравнительной степени. Далее могут следовать запросы по признакам «атрибут + значение» и «часть речи + атрибут + значение».

Также обозначим в этом блоке запрос для нахождения всех сочетаний «прилагательное + существительное» для заданного существительного, что может быть полезным для анализа качественных характеристик определенных объектов художественной реальности текста.

Говоря об исполнении (модели) графического интерфейса пользователя корпусного менеджера, выдвинем ряд технических требований. Во-первых, интерфейс должен быть интуитивно понятным. Во-вторых, он не должен быть перегружен виджетами, поэтому один и тот же виджет должен участвовать в построении нескольких запросов, особенно при сочетании нескольких критериев.

В техническом плане пользователь должен ввести нужные данные в поля ввода, выставить определенные настройки с помощью флажков, нажать кнопку активации запроса и получить вывод в специальную текстовую область, которая должна иметь опцию сохранения содержимого в текстовый файл.

По нашему мнению, графический интерфейс пользователя, который позволил бы реализовать указанные выше запросы, может быть представлен следующим образом (см. Рис. 2).

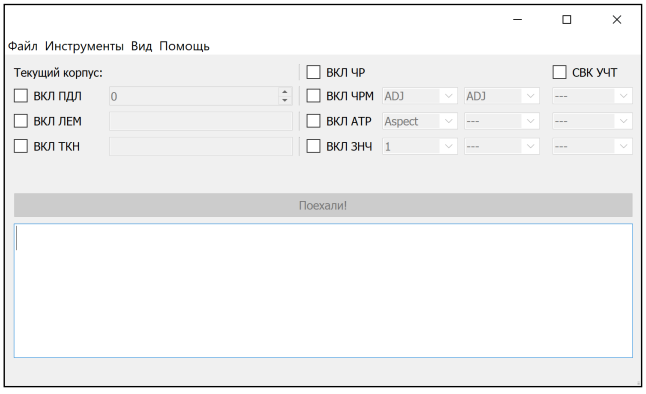

Рисунок 2. *Модель графического интерфейса пользователя корпусного менеджера*

В приведенной модели (сверху вниз) можно выделить четыре блока. Первый блок образуют меню «Файл», «Инструменты», «Вид», «Помощь». Предполагается, что через меню «Файл» можно будет сохранять результаты

вывода программы, в «Инструментах» поместятся запросы частотных списков и генератор корпуса, «Вид» будет отвечать за очистку области вывода, а «Помощь» покажет инструкцию пользователя.

Второй блок содержит настройки запросов, которые осуществляются с помощью включения или выключения флажков, выбора пунктов ниспадающего меню, а также ручного ввода данных. Здесь «ВКЛ ПДЛ» – поиск по номеру предложения, «ВКЛ ЛЕМ» – поиск по лемме, «ВКЛ ТКН» – поиск по токену, «ВКЛ ЧР» / «ВКЛ ЧРМ» – включение режима поиска по одной части речи / по многим частям речи соответственно, «ВКЛ АТР» – включение поиска по атрибутам, «ВКЛ ЗНЧ» – включение поиска по значениям атрибутов, «СВК УЧТ» – активация режима совокупного учета, то есть запроса по нескольким параметрам.

Третий блок содержит только кнопку старта запросов «Поехали!».

Четвертый блок состоит из редактируемой текстовой области вывода результатов запросов.

Далее нами была проведена оценка созданной модели, цель которой заключалась в том, чтобы проверить ее пригодность к практической реализации. Что касается написания программного кода, то удачным сочетанием, на наш взгляд, здесь является язык программирования Python и графическая библиотека PyQt, которые способны реализовать все представленные виджеты и связанные с ними функции.

Другая проблема может быть сформулирована как проблема логики построения запросов и не зависит от выбора программных средств. Иными словами, необходимо понять, удобны ли элементы управления и рациональна ли та нагрузка, которую несет каждый виджет, а также не упущены ли из вида некоторые полезные операции, которые можно производить, опираясь на содержание базы данных.

Что касается логики, то рациональным кажется наличие одной кнопки старта запроса (третий блок), хотя некоторые запросы (частотные списки) и активируются через меню «Инструменты». Каждый вид запроса приходится активировать соответствующим флажком, что немного усложняет интерфейс, но, однако, упрощает написание и чтение программного кода, поскольку нажатие кнопки «Поехали!» инициализирует проверку всех флажков по очереди, начиная с флажка «СВК УЧТ». При активации последнего программа должна проверить остальные флажки для выявления имеющихся комбинаций, например, «СВК УЧТ» + «ВКЛ ЛЕМ» + «ВКЛ ЧР» дают основание стартовать поиск по лемме с указанием ее части речи. Если выбранная комбинация не заложена в программу, то нажатие кнопки «Поехали!» ни к чему не приведет, а в области вывода должна появиться поясняющая надпись, например «Ошибка запроса» или «Такой запрос не существует».

Если пользователь нажал несколько флажков, но не выбрал «СВК УЧТ», то программа должна не выдавать ошибки, а выполнить все указанные запросы по очереди. Например, сначала вывести все заданные леммы, а затем части речи, не связывая одно с другим.

Большую функциональную нагрузку несет также поле ввода поиска по лемме, так как оно отвечает не только за поиск предложений, содержащих заданную лемму или несколько лемм, но также и за запрос для нахождения всех сочетаний «прилагательное + существительное», поскольку именно в это поле вводится искомое существительное.

Таким образом, модель предусматривает многократное использование одних и тех же виджетов, а некоторые функции вынесены в меню, что не позволяет интерфейсу выглядеть перегруженным. Единый принцип «вкл/выкл», русскоязычные подписи и защита от неверных действий оператора позволяют говорить об эргономичности интерфейса.

На завершающей стадии работы нами были написаны и протестированы мини-программы, которые позволили получить выводы – результаты прогнозируемых в модели запросов. В качестве примера приведем результаты работы мини-программы для получения частотного списка существительных (см. Табл. 1).

Таблица 1. *Результаты запроса частотного списка существительных*

Произведение	Общее количество лемм существительных	Количество уникальных лемм существительных	Первые пять единиц в списке
«На Западном фронте без перемен»	9204	2964	1. Hand : 99 2. Tag : 89 3. Auge : 81 4. Gesicht : 79 5. Kopf : 77
«Замок»	15242	3208	1. Klamm : 270 2. Herr : 214 3. Wirtin : 198 4. Gehilfe : 192 5. Schloß : 181
«1984»	19090	3358	1. time : 240 2. face : 220 3. word : 220 4. man : 187 5. moment : 169
«Затерянный мир»	13821	3063	1. man : 242 2. time : 120 3. hand : 118 4. tree : 113 5. way : 110

В последней колонке таблицы после двоеточия указано количество употреблений леммы в корпусе.

Полученные данные могут быть использованы для интерпретации содержания художественного произведения (определение смысловых доминант) и анализа идиостиля автора (в части определения лексического разнообразия текстов), в том числе в рамках сопоставительных исследований.

Заключение

Итак, в ходе нашей работы мы описали потенциальные запросы к реляционной базе данных, которые можно разделить на простые и комплексные. Далее мы построили структурную модель лингвистического корпуса и модель графического интерфейса пользователя корпусного менеджера как системы управления базой данных, каждая из которых состоит из четырех блоков.

Анализ этих моделей показал, что первая является достаточно простой и реализуемой в техническом плане, а вторая – эргономична, в том числе и за счет многократного использования одних и тех же виджетов в различных запросах, однако для повышения уровня качества исследований необходимо также предусмотреть запросы по регулярным выражениям к таблице предложений. Это позволит, например, находить контексты, содержащие токены с определенными морфемами (приставками, суффиксами и пр.). Другим полезным добавлением на этапе генерации базы данных послужит возможность ее расширения новыми текстовыми массивами, что позволит создавать динамические корпусы, необходимые, кроме прочего, для анализа дискурса СМИ.

Создание модели позволило рассмотреть потенциальный лингвистический корпус как целое, увидеть достоинства и недостатки проекта, что является важным на стадии планирования практических действий и помогает значительно сократить время разработки программного продукта.

В качестве перспективы исследования назовем непосредственно программную разработку графического интерфейса пользователя для корпусного менеджера в строгом соответствии с предложенной моделью.

Источники | References

1. Бакаев М. А., Разумникова О. М. Определение сложности задач для зрительно-пространственной памяти и пропускной способности человека-оператора // Управление большими системами: сборник трудов. 2017. № 70.
2. Бойко В. А., Легалов А. И., Зыков С. В. Архитектура интеллектуальной системы тестирования // Журнал Сибирского федерального университета. Серия «Техника и технологии». 2022. Т. 15. № 2. DOI: 10.17516/ 1999-494X-0390
3. Горожанов А. И. Экспериментальное моделирование базы данных сбалансированного лингвистического корпуса // Филологические науки. Вопросы теории и практики. 2022. Т. 15. Вып. 10. DOI: 10.30853/ phil20220563
4. Горожанов А. И., Степанова Д. В. Составление сбалансированного корпуса художественного произведения (на материале романов Ф. Кафки) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2022. № 7 (862). DOI: 10.52070/2542-2197_2022_7_862_31
5. Писарик О. И. Принципы разработки базы данных подъязыка предметной области «Строительство» // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2021. № 5 (847). DOI: 10.52070/2542-2197_2021_5_847_150
6. Читалов Д. И. Доработка графического интерфейса платформы OpenFOAM в части расширения перечня утилит для работы с расчетными сетками // Системы и средства информатики. 2022. Т. 32. № 1. DOI: 10.14357/ 08696527220113
7. Fonseca C. A., Guelpeli M. V. C., De Souza Netto R. S. Representation of structured data of the text genre as a technique for automatic text processing // Texto Livre. 2021. Vol. 15. DOI: 10.35699/1983-3652.2022.35445
8. Malyuga E. N., McCarthy M. “No” and “net” as response tokens in English and Russian business discourse: In search of a functional equivalence // Russian Journal of Linguistics. 2021. Vol. 25 (2). DOI: 10.22363/2687-0088-2021-25-2-391-416
9. O’Neill H., Welsh A., Smith D. A., Roe G., Terras M. Text mining mill: Computationally detecting influence in the writings of John Stuart Mill from library records // Digital Scholarship in the Humanities. 2021. Vol. 36 (4). DOI: 10.1093/llc/fqab010
10. Tsujii J. Natural language processing and computational linguistics // Computational Linguistics. 2021. Vol. 47 (4). DOI: 10.1162/COLI_a_00420

Информация об авторах | Author information

RU **Горожанов Алексей Иванович**[1], д. филол. н., доц.
[1] Московский государственный лингвистический университет

EN **Gorozhanov Alexey Ivanovich**[1], Dr
[1] Moscow State Linguistic University

[1] *a_gorozhanov@mail.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 17.03.2023; опубликовано (published): 17.05.2023.

Ключевые слова (keywords): моделирование; корпусная лингвистика; корпусный менеджер; графический интерфейс пользователя; spaCy; modelling; corpus linguistics; corpus manager; graphical user interface.