УДК 811.112.2’42

Горожанов А.И., д. филол. н.
(г. Москва, Россия)
ФГБОУ ВО МГЛУ
профессор кафедры грамматики и истории немецкого языка

Gorozhanov A.I., Dr of Sc. (Philology)
(Moscow, Russia)
Moscow State Linguistic University
Professor in the Department of German Language Grammar and History
e-mail: a_gorozhanov@mail.ru

ДИНАМИЧЕСКИЙ ЛИНГВИСТИЧЕСКИЙ КОРПУС КАК ИНСТРУМЕНТ КОМПАРАТИВНОГО ИССЛЕДОВАНИЯ (на материале текстов немецкоязычных СМИ)

Описываются предварительные результаты корпусного исследования на материале текстов электронных немецкоязычных изданий Spiegel и FAZ. Приводится пример анализа падежных характеристик собранных текстов. Формулируются выводы о перспективах дальнейшей работы.

Ключевые слова: динамический лингвистический корпус; немецкий язык; обработка естественного языка; Spiegel; Frankfurter Allgemeine Zeitung.

DYNAMIC LINGUISTIC CORPUS AS A COMPARATIVE RESEARCH TOOL (Based on the Texts of German-Language Media)

The preliminary results of a corpus research based on the material of texts from electronic German-language publications in Spiegel and FAZ are described. An example for the case characteristics analysis of the collected texts is given. Conclusions are formulated about the prospects for future work.

Key words: Dynamic linguistic corpus; German language; natural language processing; Spiegel; Frankfurter Allgemeine Zeitung.

В марте 2023 г. в лаборатории фундаментальных и прикладных проблем виртуального образования ФГБОУ ВО МГЛУ было инициировано очередное корпусное исследование, целью которого стало получение комплексных характеристик текстов различных изданий, а также апробация авторского метода генерации динамического лингвистического корпуса с высокой долей автоматизации процессов, включая накопление «сырого» текстового материала, его трансформацию в лингвистический корпус и получение данных, в рамках которого человек выполняет операционно-контролирующую и экспертную функции.

В предлагаемой статье кратко приводятся результаты первой фазы исследования, в ходе которой в автоматическом режиме был собран

аутентичный текстовый материал, получены два лингвистических корпуса, которые мы можем охарактеризовать как письменные одноязычные (немецкий язык) литературные публицистические исследовательские динамические размеченные сбалансированные и синхронические [1, с. 205].

Актуальность работы обусловлена необходимостью совершенствования методов корпусной лингвистики, направленных на активное привлечение библиотек обработки естественного языка в качестве программных инструментов исследования, а также на создание шаблонных решений генерации (динамических) лингвистических корпусов, которые могут быть перенесены на тексты различных стилей и жанров, и в целом – потребностью привлечения точных измерений в лингвистическую науку.

В качестве методов обозначим автоматический (программный) и автоматизированный (программный с частичной мысленной интерпретацией результата) анализ текстового материала, статистический анализ (получение формальных количественных данных), компаративный анализ (языковых средств двух корпусов), а также синтез (в части формулирования выводов).

Материалом исследования явились тексты актуальных репортажей электронных версий журнала Spiegel и газеты Frankfurter Allgemeine Zeitung (FAZ). Объем материалов по каждому изданию составляет в настоящее время около миллиона словоформ, или токенов. Тексты были собраны и обработаны в период с мая по июль 2023 г. Ведущий тип разметки – морфологическая.

После сборки обоих лингвистических корпусов посредством специализированной авторской программы [2, с. 64] нами был произведен компаративный анализ их смыслового (условно «неформального») и лингвистического (условно «формального») содержания. Пример результата последнего мы приведем ниже.

Рассмотрим распределение токенов каждого корпуса по критерию грамматической категории падежа.

Для этого в специализированном корпусном менеджере, разработанном в лаборатории фундаментальных и прикладных проблем виртуального образования, реализуем ряд запросов типа:

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

В результате такого запроса будет получено количество токенов, которым присуща характеристика «родительный падеж». Произведя запросы для каждого падежа, получим следующее (см. Таблицу 1):

Таблица 1. Распределение токенов по падежным характеристикам

Падеж	**Spiegel (%)**	**FAZ (%)**
Nom	17,52	17,28
Gen	4,43	4,43
Dat	13,68	13,38
Akk	10,7	11,5

Данные представлены в процентах относительно общего количества токенов в соответствующем корпусе.

В итоге мы получаем почти полное совпадение параметров по всем падежам. Таким образом, подъязык Шпигеля и подъязык Франкфуртер Альгемайне являются идентичными по этому критерию. Ни один из них не проявляет «архаизма», который мог бы быть заметен, например, в более высокой доле употребления родительного падежа относительно дательного.

В будущем планируется, во-первых, увеличить объемы каждого лингвистического корпуса; во-вторых, рассмотреть дальнейшие параметры для анализа их формального и неформального содержания; в-третьих, рассмотреть возможности более глубокой, при необходимости – частично ручной, разметки текстовых массивов.

Более далекая перспектива открывается в масштабировании создаваемого инструмента на различные подъязыки для создания широко востребованных учебных (см. [3]) и исследовательских электронных ресурсов.

СПИСОК ЛИТЕРАТУРЫ

1. *Горожанов А. И.*, *Степанова Д. В.* Интерпретация художественного произведения: корпусный подход // Филологические науки. Вопросы теории и практики. 2022. Т. 15, № 1. С. 203–208. DOI: 10.30853/phil20220020. EDN TCZLAF.

2. *Горожанов А. И.*, *Шевцова В. А.* Технология определения цветовой характеристики текста художественного произведения (на материале немецкого языка) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2023. № 6 (874). С. 63–68. DOI: 10.52070/2542-2197_2023_6_874_63. EDN VOUQHW.

3. *Писарик О. И.* Принципы разработки базы данных подъязыка предметной области "строительство" // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2021. № 5 (847). С. 150–160. DOI: 10.52070/2542-2197_2021_5_847_150. EDN RKPNSU.