Сравнительно-сопоставительный анализ лингвистических ресурсов для проведения корпусного анализа текстов

А. В. Дмитриев, Е. С. Крупнова

Comparative-Contrastive Analysis of Linguistic Resources for Corpus Analysis of Texts

A. V. Dmitrijev, E. S. Krupnova

Александр Владиславович Дмитриев – кандидат филологических наук, доцент; Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

E-mail: avd84@list.ru

Елена Сергеевна Крупнова – магистр, специалист по учебно-методической работе; Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

E-mail: krupnalena@mail.ru

Alexander V. Dmitrijev – Candidate of Philological Sciences, Associate Professor; Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russian Federation

ORCID: 0000-0003-3632-793X

Elena S. Krupnova – Master's degree, specialist in educational and methodological work; Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russian Federation

ORCID: 0009-0007-3127-2737

Статья поступила: 05.10.2024. Принята к печати: 20.10.2024.

Received: 05.10.2024. Accepted for publication: 20.10.2024.

В статье рассматривается основная задача корпусной лингвистики – корпусный анализ письменных текстов на естественном языке с помощью лингвистических ресурсов, которые используются для её решения. Корпусный анализ подразумевает метод исследования языка, который использует большие коллекции текстов или корпуса для получения статистических и лингвистических данных о языке. Лингвистические ресурсы, такие как словари, тезаурусы, грамматические базы данных значительно расширяют возможности и точность корпусного анализа. Помимо этого, корпусная лингвистика занимается созданием корпусных менеджеров, которые обрабатывают тексты и выполняют функции составления конкорданса, поиска ключевых слов, коллокаций и другие. В работе кратко описывается функционал программ WMatrix, WordSmith, GATE, AntConc и Sketch Engine, а также проводится сравнительно-сопоставительный анализ их характеристик. В результате сделан вывод о том, что ряд программ отличается набором функций, параметрами сохранения данных, форматом входного текста и доступностью. Кроме того, перечисляются направления их использования в научно-практической деятельности. Лингвистические ресурсы могут быть полезны для стилистического анализа текстов, изучения лингвистических особенностей авторского стиля, обучения иностранному языку, например, грамматике или лексике, в компьютерной лексикографии, дискурс-анализе и в других направлениях. Рассмотренные инструменты не только повышают точность анализа, но и расширяют возможности, интегрируясь в программные инструменты для автоматизации корпусного анализа. Выбор подходящего инструмента для проведения исследования зависит от объёма и глубины анализа текста.

In the last few decades, a scientific field known as computational linguistics has been actively developing. The paper discusses the main task of corpus linguistics – corpus analysis of written natural-language texts with the help of linguistic resources that are used to solve it. Corpus analysis refers to a method of language research that utilizes large collections of texts or corpora to obtain statistical and linguistic data about the language. Linguistic resources such as dictionaries, thesauri, and grammatical databases greatly enhance the capability and accuracy of corpus analysis. In addition, corpus linguistics deals with the building of corpus managers that process texts, perform concordance, search for keywords and collocations, etc. The paper briefly describes the functionality of WMatrix, WordSmith, GATE, AntConc and Sketch Engine programs and makes a comparative-contrastive analysis of their characteristics. It is concluded that the programs differ in feature set, data saving parameters, input text format and accessibility. In addition, directions for their use in research and practice are suggested. Linguistic resources can be useful for stylistic analysis of texts, studying linguistic features of author's style, teaching a foreign language, for example, grammar or vocabulary, in computer lexicography, discourse analysis and other directions. The example of the corpus analysis of the topic *famine* during the blockade of Leningrad with the help of the AntConc program is given. In the course of the mentioned research, 749 fragments of memories of Leningrad citizens were collected on the basis of 15 frequency words and a frequency dictionary of 158 words was compiled. Considered tools not only increase the accuracy of analysis, but also expand the possibilities and integrate into software tools for automation of corpus analysis. The choice of the appropriate tool for the study depends on the scope and depth of text analysis.

Ключевые слова: корпусная лингвистика, лингвистические корпуса, корпусный менеджер, стилистический анализ текста

Keywords: natural language processing, corpus linguistics, linguistic corpora, corpus manager, corpus stylistics, stylistic corpus analysis

УДК 81’322:81`42

OECD: 6.02OT

Постановка проблемы. В последние несколько десятков лет активно развивается научная область, компьютерная лингвистика, основным фокусом которой является автоматическая обработка письменных текстов на естественном языке (Natural Language Processing или NLP). Компьютерная лингвистика – междисциплинарная область, которая возникла на стыке лингвистики, информатики, математики и искусственного интеллекта. Основными её прикладными задачами являются машинный перевод, классификация и кластеризация текстов, поиск и извлечение информации, индексирование, реферирование, аннотирование, интеллектуальный анализ данных, формирование ответов на вопросы, анализ тональности текстов, распознавание и синтез звучащей речи и другие.

Анализ языковых данных является непростой задачей и для её решения требуется большой массив данных, в котором содержится несколько сотен тысяч примеров употреблений тех или иных слов. Этим вопросом занимается другой подраздел компьютерной лингвистики – корпусная лингвистика, которая занимается созданием и использованием лингвистических корпусов для решения различных задач в области лингвистики и смежных областей [Николаев, 2016]. Актуальность исследования заключается в необходимости автоматизации процесса обработки текстов на естественном языке и их корпусном анализе для решения различных прикладных задач, как поиск и извлечение информации, анализ кореферентности, машинное обучение, обучение переводу и многие другие. На данный момент существует множество инструментов, с помощью которых можно проанализировать тексты, однако среди них можно выявить ряд отличий в наборе функций и инструментов. Для получения наиболее эффективных результатов проведённого исследования необходимо знать, какой лингвистический ресурс подойдёт лучше всего. Кроме того, программы со временем дорабатываются и в них появляются новые функции, которые не отражены в некоторых существующих работах.

История вопроса. Вопросы использования корпусов и корпусных менеджеров были рассмотрены в [Захаров, 2005]. Лингвистический корпус – «большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [Захаров, 2005, с. 3]. Термин «корпус текстов» также обозначается «корпусным менеджером», специализированной системой поиска, с помощью которой можно искать ключевые слова, фразы, коллокации и контексты словоформ в корпусе, создавать конкордансы, составлять списки слов по заданным критериям, получать статистическую информацию о частоте употребления слова в корпусе и представлять результаты в удобном формате для пользователя. Такие программы быстро обрабатывают результаты и интуитивно понятны в использовании.

Рассматриваемые в статье программы были проанализированы с точки зрения функционала в некоторых работах. Например, сравнение двух программ Antconc и Sketch Engine для исследования коллокаций в английском языке, отдельное описание функционала программы Antconc, Sketch Engine, а также сравнение Antconc, Wordsmith Tools и Sketch Engine на материале текстов кинодискурса [Палийчук, 2022; Котюрова, 2020; Кротова, 2019; Шамова, 2021].

Методология и методика исследования. Цель исследования заключается в сравнительно-сопоставительном анализе пяти наиболее популярных лингвистических инструментов – WMatrix, WordSmith, GATE, AntConc и Sketch Engine, используемых для автоматической обработки и корпусного анализа текстов, и их применении в научно-практической деятельности.

Исследование проводится с применением следующих методов: отбор и анализ данных о функциональности наиболее популярных лингвистических ресурсов, сравнительно-сопоставительный анализ инструментов и метод комплексного описания полученных результатов.

Анализ материала. Для проведения анализа были взяты 5 наиболее популярных лингвистических ресурса WMatrix, WordSmith, GATE, AntConc and Sketch Engine. Кратко охарактеризуем программы.

WMatrix. Программа была разработана в конце XX века Полом Рейсоном в рамках проекта REVERE, цель которого состояла в изучении вопроса извлечения информации из документов, связанных с разработкой программного обеспечения. С помощью данной программы можно выделять ключевые слова, определять часть речи слова, проводить анализ текстов на уровне грамматики и семантики, визуализировать частотность употребления слов в корпусе и анализировать конкордансы, исследуя единицы языка в их контекстуальном окружении. Простой функционал программы WMatrix6 состоит из четырёх функций: поиск списка слов и их частотность, поиск конкретного слова и облака слов, позволяющие увидеть значимость слов [Rayson]. Стоит также отметить, что на данный момент доступ к программе ограничен.

После загрузки текста в программу можно провести анализ данных: увидеть список наиболее частотных слов по частям речи и по семантическому тегу; класс слов, отражающих тему текста; конкорданс слова или тега и просмотр контекста слева или справа; визуализацию частотности слов и тегов (чем крупнее шрифт, тем наиболее значимо слово в тексте или, другими словами, тем чаще оно появляется в тексте). Кроме того, есть возможность сравнить файл со стандартным эталонным корпусом для нахождения ключевых слов и ключевых семантических категорий, связанных с эталонным корпусом.

В WMatrix5 можно также провести более продвинутый анализ данных как токенизацию, получение многословных выражений (семантический теггер автоматически размечает их как единое целое с помощью одного семантического тега) и n-грамм (или их ещё называют кластерами, лексическими связками или устойчивыми выражениями). Кроме того, доступны функции получения списка частотности лемм, частотности тега по части речи и анализа ключевых частей речи, а также коллокаций (словосочетания, которые встречаются либо слева, либо справа от

слова). Ещё одной особенностью программы является анализ метафор с использованием семантических тегов.

Ещё одним инструментом являются коллокации. В программе в колонке “Word” можно увидеть слово, а в “Collocate” словосочетание, которое встречается либо слева, либо справа от него. В других шести колонках (L3 – R3) содержится информация о частотности появления словосочетания в позиции слева от слова – одно слово (L1), два слова (L2) и три слова (L3). Таким же образом в позиции справа от слова – R1, R2 и R3. В колонке “Total” указывается сумма всех позиций, в которых словосочетание встречается со словом. В колонке “Word Freq” – число встречаемости слова в корпусе, в “Collocate Freq” – число словосочетаний. На основе этой информации можно составить таблицу для каждого слова и пары слов и посчитать показатели статистических данных. Двумя основными метриками в программе являются MI (взаимная информация) и LL2 (the two-cell Log-Likelihood или правдоподобие). С помощью первой можно выделить устойчивые коллокации.

Sketch Engine. Проект, разработанный лингвистом Адамом Киллгарриффом и чешским программистом Павлом Рыхли. С помощью этого инструмента можно анализировать аутентичные тексты, для выявления необычных в использовании и редких элементов языка. Он также предназначен для анализа текста или приложений для интеллектуального анализа текста.

Программа обеспечивает репрезентативную выборку языка, поскольку содержит 800 готовых к использованию корпусов на более чем 100 языках, каждый из которых имеет размер до 80 миллиардов слов [программа Sketch Engine].

Sketch Engine предоставляет большой набор функций, основными которого являются: Concordance (поиск примера употребления вводимого слова, леммы, фразы или словосочетания); Word list (список всех слов в выбранном корпусе); Keywords and Terms (ключевые слова и термины); Collocations (коллокации, стоящие слева или справа от заданного слова); Thesaurus (нахождение слов, которые появляются в похожем контексте, как и вводимое слово); Word Sketch (описания возможной сочетаемости слова с другими); Word Sketch Differences (сравнение скетчей для двух лексических единиц); WebBootCaT (создание собственного корпуса текстов) и Trends (изменение частоты слов в корпусе) [Палийчук, 2022].

Лингвистический процессор GATE (General Architecture for Text Engineering) предназначен для решения различных задач: ручная и автоматическая семантическая разметка текстов, извлечение информации, анализ кореферентных связей в тексте, машинное обучение, работа с онтологиями [Рубайло, 2016]. Последняя выпущенная версия на 2024 год – версия 9.0.1.

Первая версия системы GATE была выпущена в 1996 году разработчиками из университета Шеффилда. Она был написана на языке C++ и использовалась в широком спектре контекстов анализа языка, включая извлечение информации на английском, греческом, испанском, шведском, немецком, итальянском и французском языках [Большакова, 2011].

Данная система включает в себя три важные составляющие: архитектура, фреймворк и среда разработки. Под архитектурой понимается абстрактное описание того, как может быть построена система обработки языка, типы обычно используемых

компонентов и так далее. Компоненты – многократно используемые программные блоки с четко определенными интерфейсами. Под фреймворком подразумевается объектно-ориентированная библиотека классов, реализующая архитектуру и предоставляющую ряд сервисов, которые можно использовать в различных контекстах приложений. Одним из таких приложений является среда разработки, построенная на основе фреймворка.

Платформа GATE поддерживает множество текстовых форматов, включая Plain Text, PDF, Email, форматы Microsoft Office и другие. В программе есть возможность сохранить корпуса текстов и документы и продолжить с ними работать в дальнейшем.

Преимуществами данной программы являются следующие:

- удобный графический интерфейс;
- экспорт аннотаций в XLM формате;
- возможность повторного использования лингвистических компонентов;
- представление общей базы для разработки приложений и компонентов.

Недостатками являются:

- сложность и неочевидность графического интерфейса;
- отсутствие возможности добавления новых компонентов для визуализации данных;
- неэффективная реализация базы данных;
- необходимость совместимости схем аннотаций для интеграции различных компонентов;
- отсутствие поддержки компонентов, представляющих источники данных [Большакова, 2011].

WordSmith Tools. Программный пакет, предназначенный в первую очередь для работы с текстовыми файлами. Набор программ был разработан британским лингвистом Майком Скоттом и впервые выпущен в 1996 году. На данный момент доступна бесплатная версия 9.0. Стоит отметить, что WordSmith Tools можно использовать на 80 языках.

Программный пакет состоит из основных трёх модулей:

1. Модуль “Concord” используется для создания конкордансов (списка всех употреблений вводимого слова или фразы в контексте). С его помощью можно отобрать коллокации или кластеры слов, а также их местоположение в тексте.
2. Модуль “WordList” используется для составления списка всех слов или словоформ, включённых в выбранный корпус, а также частотного списка.
3. Модуль “KeyWord” используется для создания списка ключевых слов и грамматических форм в соответствии с определёнными статистическими критериями.

Кроме этого, каждый модуль содержит в себе другие функции анализа текста, как, например, поиск словосочетаний – коллекции слов, которые наиболее часто используются в тексте вместе с одним определенным словом. Также есть ряд

дополнительных модулей, которые полезны для подготовки, очистки и форматирования текстового корпуса. Главным преимуществом программы является выгрузка результатов на экран или их сохранение в отдельный файл.

AntConc. Бесплатный программный инструмент, разработанный Лоуренсом Энтони в 2011 году, для проведения статистических исследований текстов. С его помощью можно идентифицировать и подсчитывать длинные кластеры, составлять конкорданс для заданного слова в пределах контекстного окна, частотный список словоформ или лемм с указанием ранга и абсолютной частоты; выделять ключевые слова художественного текста; выявлять связи между полученным ключевым словом и анализом ключевой семантической области; искать длинные n-граммы для определения их значимости, ценности и связи со значением слова [Николаев, 2016]. Помимо этого, данная программа используется для стилистического анализа текста на основе методов корпусной лингвистики. AntConc выполняет следующие функции:

1) построение конкорданса для заданного слова в пределах контекстного окна;
2) построение графика к конкордансу;
3) построение частотного списка словоформ или лемм с указанием ранга и абсолютной частоты
4) выделение ключевых слов художественного текста, которые могут помочь читателям понять смысл текста.
5) выявление связи между полученным ключевым словом и анализом ключевой семантической области.
6) поиск коллокаций заданного слова на основе мер ассоциации;
7) поиск длинных n-грамм для определения их значимости, ценности и связи со значением слова [Николаев, 2016].

Программа состоит из девяти инструментов: KWIC (результаты поиска употребления слов или фраз в корпусе текстов); plot (график конкорданса) для представления результатов поиска в виде штрих-кода, который показывает местоположение, где слово или фраза появляется в корпусе; функция File (просмотр файла) отображает содержание текста; clusters (кластеры) показывает группы слов, которые появляются в тексте рядом с ключевым словом, N-Gram (списки n-gram) для отображения наиболее частотных словосочетаний в корпусе; функция Collocate для нахождения слов, которые сочетаются с другими словами; Word list для получения упорядоченного списка всех слов в корпусе от наиболее частотных к менее, Keyword для определения слов, часто употребляемых в текстах по сравнению с эталонным корпусом, и Wordcloud (визуализация результатов, полученных с помощью инструментов KWIC, File, Cluster, N-Gram, Collocate, Word и Keyword, в виде «облака слов» (например, наиболее частотные слова представлены шрифтом большего размера). Ограничением данной программы заключается в объёме корпуса – допускается небольшой размер.

Сравним в таблице представленные выше программы по функционалу, формату вводимого текста/файла, доступности, сохранении данных и языку интерфейса.

Таблица 1. Сравнительный анализ лингвистический ресурсов

	WMatrix	**AntConc**	**WordSmith Tools**	**GATE**	**Sketch Engine**
конкорданс	+	+	+	–	+
график конкорданса	–	+	–	–	–
кластеры	–	+	+	–	+
N-граммы	+	+	–	–	+
коллокации	+	+	+	–	+
частотность слов	+	+	+	–	+
список слов	+	+	+	–	+
ключевые слова	+	+	+	–	+
Wordcloud	+	+	–	–	+
разметка текста	+	–	+	+	+
формат файла/ текста	plain text format, html, sgml, xml	txt, srt, csv, tsv, html, xml, docx, pdf	plain text format, html, xml	plain text format, html, sgml, xml, rtf, Email, OpenOffice, pdf, Microsoft Office	txt, doc, docx, pdf, xml, html, htm, zip, tar.bz2, tar.gz, tgz, vert, ps, tmx
доступность	–	+	+	+	+ (30 дней бесплатно)
сохранение данных	–	–	+	+	+
интерфейс на русском языке	–	–	–	–	–

Результаты исследования

Исходя из результатов составленной таблицы, можно сделать следующие выводы:

1. AntConc, Sketch Engine, WMatrix и WordSmith позволяют работать с коллокациями, кластерами, ключевыми словами текстов, n-граммами, списками слов и конкордансом. Их функционал похож и отличается несколькими параметрами.
2. В программах Antconc, GATE и Sketch Engine файлы могут быть загружены в большом ряде форматов, что удобно в их использовании.
3. Стоит также отметить, что из всех представленных программ только WMatrix ограничен в доступе на момент 2024 года. Остальные бесплатные и их можно скачать локально.
4. В программах WordSmith Tools, GATE и Sketch Engine есть возможность сохранять полученные данные и использовать их повторно.
5. Интерфейс всех ресурсов доступен только на английском языке, однако он интуитивно понятен в использовании.

В целом можно отметить, что для подробного стилистического анализа текста наиболее подходящей является программа Sketch Engine или AntConc. Первая программа имеет такой недостаток, как ограниченный период бесплатного доступа, а вторая — не предусматривает функцию сохранения полученных данных, что может быть неудобным и требует переноса результатов в документ.

В связи с тем, что рассмотренные лингвистические ресурсы по обработке и корпусному анализу текстов были созданы несколько лет назад, они были использованы в различных направлениях, как изучение онтологий, обучение чат-ботов на основе сравнения диалогов между людьми и человеком и компьютером, сравнение письменного и устного типов речи английского языке, анализ ключевых слов и тем в литературных текстах, фразеология, метафоры в политическом или бизнес дискурсе, анализ интернет-блогов Синглиша, изучение языковых вариаций с помощью метода извлечения ключевых тем, использование корпусов в процессе обучения специальному переводу и другие. Приведём несколько направлений использования корпусных менеджеров и систем в научной и практической деятельности.

Одним из актуальных направлений является корпусная стилистика – область, которая стала активно развиваться с конца 1960-х годов, но популярным стала в 2000-е годы. С 2007 года отмечается «корпусный поворот» в стилистике, о чём свидетельствует большое количество энциклопедических статей о практике проведения анализа, а также монографий, в которых описывается данный подход [Leech, 2007; Mahlberg, 2012b; McIntyre, 2015].

Помимо изучения стиля текста при анализе слов, фраз и предложений, частей речи, знаков препинания, пауз, выявления языковых особенностей, которые остаются незамеченными при ручном анализе текста, программы используются для стилистического анализа художественных текстов, анализа фразеологии, сегментации текста, связности и когерентности [Fischer, 2010; Mahlberg, 2012a].

Лингвистические инструменты могут быть также использованы для исследования парцелляции в газетном дискурсе, английской абсолютной причастной конструкции, морфологических и грамматических ошибок в письменной речи носителей языка, дискредитирующих тактик в дискурсе социальных медиа и других работах.

В качестве ещё одного направления исследования можно рассмотреть стилистический анализ текстов с применением метода корпусной лингвистики. Например, анализ темы «голод» во время блокады Ленинграда при рассмотрении и описании жизни жителей города, описанная в книге Алеся Адамовича и Даниила Гранина «Блокадная книга». Для проведения корпусного анализа была выбрана программа AntConc и использованы её четыре функции: Wordlist, Concordance, File view и Clusters.

Из полученного списка слов были отобраны 15 по теме «голод» и взяты их основы: голод*, хлеб*, холод*, смерт*, грамм*, карточк*, ужас*, кусоч*, труп*, съе*, испытани*, еда, слаб*, дистрофи* и болезн* (звёздочкой обозначено любое количество символов после основы). Далее были найдены контексты употребления отобранных слов. В результате исследования было собрано 747 отрывков. Большее количество употребления основ слов «хлеб*», «голод*», «смерт*», «карточк*» и «съе*». Кроме того, был составлен частотный тематический словарь, в который вошли 158 словосочетаний. Наиболее частотными оказались следующие: граммов хлеба, от голода, по карточкам, за хлебом и на хлеб [Крупнова, 2024].

Ещё одним возможным применением лингвистического ресурса является отбор лексических единиц для тематического словаря предметной области для системы фразеологического машинного перевода. Для начала необходимо создать корпус рефератов, например, собрать их с сайта базы данных ВИНИТИ РАН путем парсинга, указав тематику и запрос. Затем загрузить файл в формате txt и провести анализ корпуса с помощью программы AntConc. В первую очередь отобрать самые частотные значимые существительные, затем провести анализ кластеров с этими словами, которые могут войти в словарь; анализ N-gram и употребление ключевых слов или фраз в контексте.

Корпусные менеджеры также можно использовать на уроках иностранного языка при изучении, например, грамматики и лексики. С их помощью можно познакомить учащихся с тенденциями употребления тех или иных конструкций, а также слов у письменной и устной речи. Кроме того, для сравнения можно взять тексты разных стилей и жанров.

Выводы. В последние несколько лет активно развивается область на стыке нескольких дисциплин, компьютерная лингвистика, в задачи которой входит обработка текстов на естественном языке. Для автоматизации данного процесса и получения быстрого и точного результата разрабатываются лингвистические ресурсы, которые анализируют язык на разных уровнях.

Корпусная лингвистика занимается созданием корпусных менеджеров, которые обрабатывают тексты и выполняют функции составления конкорданса, поиска ключевых слов, коллокаций и другие. В данной работе были рассмотрены

шесть наиболее популярных программы, среди которых выделяются WMatrix, AntConc, WordSmith Tools, GATE и Sketch Engline.

В ходе сравнительного анализа шести наиболее популярных программ были сформулированы выводы о том, что WordSmith Tools и AntConc предоставляют возможность работы с коллокациями, кластерами, ключевыми словами текстов, а также списками слов. Среди всех ресурсов WMatrix является единственной недоступной. В целом, в ряде программ есть ограниченный набор инструментов, отсутствие возможности сохранения данных, а также отсутствие интерфейса на русском языке.

В работе были также сформулированы некоторые возможные пути их использования в научной и практической деятельности. Среди актуальных направлений развития сетевых лингвистических ресурсов можно выделить корпусную стилистику; машинный перевод, компьютерную лексикографию, дискурс анализ и использование в образовательном пространстве.

Таким образом, лингвистические ресурсы в области автоматической обработки и корпусного анализа текста могут быть полезны при решении различных задач. Выбор подходящего инструмента зависит от объёма и глубины анализа текста.

Благодарности: исследование профинансировано Министерством науки и высшего образования РФ в рамках Программы стратегического академического лидерства «Приоритет-2030» (соглашение No 075-15-2024-201 от 6 февраля 2024 г.).

Литература

Большакова, Е. И., Клышинский, Э. С., Ландэ, Д. В., Носков, А. А., Пескова, О. В., Ягунова, Е. В (2011) Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. Москва: МИЭМ.

Захаров, В. П. (2005) Корпусная лингвистика: учебно-методическое пособие. Санкт-Петербург: Изд-во СПбГУ.

Котюрова, И. А. (2020) Корпусные исследования с помощью сервиса Antconc в условиях работы в вузе. *Язык и культура,* 52, 36–50. DOI: 10.17223/19996195/52/3

Кротова, Е. Б. (2019) Sketch Engine для лингвистических исследований. *Германистика сегодня: материалы Международной практической конференции, 16-17 октября 2018 г., Казань.* Казань: Изд-во Казан. ун-та. 107–112.

Крупнова, Е. С. (2024) Корпусный анализ темы «голод» во время блокады Ленинграда и составление частотного словаря. *Второй Международный молодёжный конкурс научных проектов «Стираем границы»: сборник материалов.* Москва: РГУ им. А. Н. Косыгина. 143–146.

Николаев, И. С., Митренина, О. В., Ландо, Т. М. (редакторы) (2016) Прикладная и компьютерная лингвистика: коллективная монография. 2-е изд. Москва: ЛЕЛАНД.

Палийчук, Д. А. (2022) Корпусные технологии в изучении коллокаций (на примере сервисов «AntConc» И «SketchEngine»). *Studia Humanitatis*, 2, 13–14. URL: https://cyberleninka.ru/article/n/korpusnye-tehnologii-v-izuchenii-kollokatsiy-na-primere-servisov-antconc-i-sketchengine

AntConc: бесплатный набор инструментов для корпусного анализа, позволяющий конкордансировать и анализировать текст // Сайт Лоуренса Энтони: официальный сайт. URL: https://www.laurenceanthony.net/software/antconc/

Программа Sketch Engine // Sketch Engine: официальный сайт. URL: https://www.sketchengine.eu/

Рубайло, А. В., Косенко, М. Ю. (2016) Программные средства извлечения информации из текстов на естественном языке. *Альманах современной науки и образования, 12* (114), 87–92.

Шамова, Н. А. (2021) Сравнительно-сопоставительный анализ корпусных инструментов (на примере работы с корпусами кинодискурса). *Вестник Нижегородского государственного лингвистического университета им. Н. А. Добролюбова,* 53, 82–95. DOI: 10.47388/2072-3490/lunn2021-53-1-82-95

Fischer-Starcke B. (2010) Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries. London; New York: Continuum.

Leech, G., Short, M. (2007) Style in Fiction: A Linguistic Introduction to English Fictional.

London; New York: Longman. URL: https://sv-etc.nl/styleinfiction.pdf

Mahlberg M. (2012). Corpus Stylistics and Dickens's Fiction. New York: Routledge

Mahlberg, M. (2012). The corpus stylistic analysis of fiction – or the fiction of corpus stylistics? *Corpus Linguistics and Variation*

References

AntConc program: official website. Retrieved from https://www.laurenceanthony.net/software/antconc/

Bolshakova, E. I., Klyshinsky, E. S. (2011). Automatic processing of texts in natural language and computational linguistics: textbook. Moscow: Moscow Institute of Electronics and Mathematics Publ., 272. (In Russian).

Fischer-Starcke, B. (2010). Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries. London; New York: Continuum.

Kotyurova, I. A. (2020). Corpus-based studies with Antconc service at the university. *Language and Culture,* 52, 36–50, 10.17223/19996195/52/3. (In Russian).

Krotova, E. B. (2019). Sketch Engine for linguistic research. *Germanistics Today: Proceedings of the International Practical Conference*. Kazan: Kazan Publ. Kazan University, 107–112. (In Russian).

Krupnova E.S. (2024). Corpus analysis of the theme "hunger" during the blockade of Leningrad and the compilation of a frequency dictionary. *Second International Youth Competition of Scientific Projects "Erasing Borders": collection of materials.* Moscow: Kosygin Russian State University Publ., 143–146. (In Russian).

Leech, G., Short, M. (2007). Style in Fiction: A Linguistic Introduction to English Fictional. London; New York: Longman. Retrieved from: https://sv-etc.nl/styleinfiction.pdf.

Mahlberg M. (2012). Corpus Stylistics and Dickens's Fiction. New York: Routledge.

Mahlberg, M. (2012). The corpus stylistic analysis of fiction – or the fiction of corpus stylistics? Corpus Linguistics and Variation in English, 75, 77–95, 10.1163/9789401207713_008.

McIntyre D. (2015). Towards an integrated corpus stylistics. *Topics in Linguistics*, 16(1), 59–69, 10.2478/topling-2015-0011. Retrieved from: http://dx.doi.org/10.2478/topling-2015-0011.

Nikolaev, I. S., Mitrenina, O. V., Lando, T. M. (2016). Applied and computational linguistics. Collective monograph. 2nd ed. Moscow: LELAND Publ. (In Russian).

Paliychuk, D. A. (2022). Corpus technologies in the study of collocations (by the example of "AntConc" and "SketchEngine" services). *Studia Humanitatis*, 2, 13–14. Retrieved from: https://cyberleninka.ru/article/n/korpusnye-tehnologii-v-izuchenii-kollokatsiy-na-primere-servisov-antconc-i-sketchengine. (In Russian).

Rayson, P. (2009). Wmatrix: a Web-based Corpus Processing Environment. Retrieved from: http://ucrel.lancs.ac.uk/wmatrix/.

Rubaylo, A. V., Kosenko, M. Yu. (2016). Program means of information extraction from natural language texts. *Almanac of Modern Science and Education,* 12(114), 87–92. (In Russian).

Shamova, N. A. (2021). Comparative-comparative analysis of corpus tools (on the example of work with film discourse corpora). *Bulletin of N.A. Dobrolyubov Nizhny Novgorod State Linguistic University*, 53, 82–95, 10.47388/2072-3490/lunn2021-53-1-82-95. (In Russian).

in English. Availability: Published, 75, 77–95. DOI: 10.1163/9789401207713_008

McIntyre, D. (2015) Towards an integrated corpus stylistics. *Topics in Linguistics,* 16 (1), 59–69. URL: https://topling.ukf.sk/index.php/topling/article/view/22/22. DOI: http://dx.doi.org/10.2478/topling-2015-0011

Rayson, P. (2009) Wmatrix: a Web-based Corpus Processing Environment. *Computing Department, Lancaster University.* URL: http://ucrel.lancs.ac.uk/wmatrix/

Wmatrix tutorial (for version 5). Documentation: Step-by-step instructions using a case study of linguistic analysis of political party manifestos for the UK General Election (updated November 2022). URL: https://ucrel.lancs.ac.uk/wmatrix/tutorial/

Wmatrix tutorial (for version 6). Documentation: Step-by-step instructions on the example of linguistic analysis of political party manifestos for the UK General Election (updated in June 2023). URL: https://ucrel.lancs.ac.uk/wmatrix/tutorial6/

Introduction to WordSmith Tools. *WordSmith site*. URL: https://lexically.net/downloads/version_64_8/HTML/index.h tml

Sketch Engine program: official website. Retrieved from: https://www.sketchengine.eu/.

WMatrix 5. Documentation: Step-by-step instructions using a case study of linguistic analysis of political party manifestos for the UK General Election (updated November 2022). Retrieved from: https://ucrel.lancs.ac.uk/wmatrix/tutorial/.

WMatrix 6. Documentation: Step-by-step instructions on the example of linguistic analysis of political party manifestos for the UK General Election (updated in June 2023). Retrieved from: https://ucrel.lancs.ac.uk/wmatrix/tutorial6/.

WordSmith Tools. Retrieved from: https://lexically.net/downloads/version_64_8/HTML/index.html.

Zakharov, V. P. (2005). Corpus linguistics: Manual. Saint Petersburg: Saint Petersburg State University Publ. (In Russian).

Для цитирования статьи:

Дмитриев, А. В., Крупнова, Е. С. (2024). Сравнительно-сопоставительный анализ лингвистических ресурсов для проведения корпусного анализа текстов. *VERBA. Северо-Западный лингвистический журнал*, 3(13), 24–35. DOI: 10.34680/VERBA-2024-3(13)-24-35

For citation:

Dmitrijev, A. V., Krupnova, E. S. (2024). Comparative-Contrastive Analysis of Linguistic Resources for Corpus Analysis of Texts. *VERBA. North-West linguistic journal*, 3(13), 24–35. (In Russian). DOI: 10.34680/VERBA-2024-3(13)-24-35