УДК 81-13
Науч. спец. 5.9.8
DOI: 10.36809/2309-9380-2025-47-84-89

Зоя Алексеевна Кратнова
Ярославский государственный педагогический университет им. К. Д. Ушинского,
старший преподаватель кафедры китайского языка, Ярославль, Россия
e-mail: zais.87@mail.ru

Автоматизированный анализ научного текста

Аннотация. Основной целью данного исследования является сравнительный анализ различных программных инструментов для обработки текстов, таких как AntConc, LancsBox, WordSmith Tools и «Текстометр». Исследование направлено на выявление их функциональных возможностей, эффективности, удобства использования и областей применения в лингвистических и прикладных исследованиях. В исследовании проведен анализ русскоязычного научного текста кандидатских и докторских диссертаций по различным специальностям с применением вышеуказанных инструментов.

Ключевые слова: программный инструмент, обработка текста, конкордансный анализ (KWIC), синтаксический анализ, диссертация, коллокация.

Zoya A. Kratnova
Yaroslavl State Pedagogical University named after K. D. Ushinsky,
Senior Lecturer of the Chinese Language Department, Yaroslavl, Russia
e-mail: zais.87@mail.ru

Automated Analysis of Scientific Text

Abstract. The main purpose of this study is a comparative analysis of various software tools for text processing, such as AntConc, LancsBox, WordSmith Tools and Textometer. The study is aimed at identifying their functionality, effectiveness, usability and areas of application in linguistic and applied research. The study analyzed the Russian-language scientific text of candidate and doctoral dissertations in various specialties using the above tools.

Keywords: software tools, analysis of the text, concordance analysis, syntactic analysis, dissertation, collocation.

Введение (Introduction)

В современном мире, где информация и данные играют ключевую роль в научных исследованиях, профессиональной деятельности и даже повседневной жизни, инструменты для анализа текстов становятся всё более востребованными. Основной целью данного исследования является сравнительный анализ различных программных инструментов для обработки текстов, таких как AntConc, LancsBox, WordSmith Tools и «Текстометр». Исследование направлено на выявление их функциональных возможностей, эффективности, удобства использования и областей применения в лингвистических и прикладных исследованиях. Это позволит определить, какой из инструментов является наиболее подходящим для решения конкретных задач, связанных, в частности, с анализом текстов научного стиля, а именно текстов диссертаций.

Теоретическая основа данного исследования строится на нескольких ключевых концепциях и теориях, связанных с компьютерной лингвистикой, корпусной лингвистикой и программной инженерией.

В данном исследовании проведен анализ русскоязычного научного текста кандидатских и докторских диссертаций по различным специальностям: 10.02.04 «Германские языки», 10.02.05 «Романские языки», 10.02.19 «Теория языка», 10.02.20 «Сравнительно-историческое, типологическое и сопоставительное языкознание», 10.02.22 «Языки народов зарубежных стран Европы, Азии, Африки», 10.02.01 «Русский язык», 5.9.5 «Русский язык. Языки народов России», 5.9.6 «Языки народов зарубежных стран (с указанием конкретного языка или группы языков)», 5.9.8 «Теоретическая, прикладная и сравнительно-сопоставительная лингвистика», — за последнее десятилетие*.

Методы (Methods)

В предпринятом исследовании используются следующие методы: метод сравнительного анализа, а именно

* Картотеку исследования составили диссертации, защищенные до и после изменений в 2021 г. номенклатуры научных специальностей, в связи с этим указаны шифры специальностей, как действовавшие до внесения изменений, так и актуальные.

Для цитирования: Кратнова З. А. Автоматизированный анализ научного текста // Вестник Омского государственного педагогического университета. Гуманитарные исследования. 2025. № 2 (47). С. 84–89. DOI: 10.36809/2309-9380-2025-47-84-89

частотный анализ, конкордансный анализ (KWIC), коллокационный анализ, синтаксический анализ, экспериментальный метод, контент-анализ.

Результаты и обсуждение (Results and Discussion)

Анализ приложения для обработки текстов LancsBox

LancsBox [1] — приложение для обработки текстов, разработанное в рамках проекта, финансируемого Ланкастерским университетом, с целью создания доступного инструмента для анализа текстов. Проект был запущен в начале 2010-х гг. и с тех пор претерпел множество изменений и улучшений на основе обратной связи от пользователей и новых научных данных.

Основная цель создания LancsBox заключалась в предоставлении инструмента, который бы сочетал аналитические возможности с простотой использования. Разработчики стремились создать приложение, которое бы позволило пользователям легко интерпретировать и анализировать большие объемы текстовых данных, будь то научные статьи, исторические документы или обычные тексты из интернета.

Одной из ключевых функций LancsBox является его способность работать с текстовыми корпусами. Благодаря корпусному анализу исследователи могут анализировать объемные тексты, проводить статистический анализ, а также выявлять лингвистические закономерности и тенденции. Приложение LancsBox включает множество инструментов для детального анализа слов и словосочетаний, что позволяет выявлять частотные конструкции и анализировать их значение и контекст.

Приложение незаменимо для создания наглядности и информативности посредством инструментов для визуализации данных. С помощью данного инструмента можно создавать графики, схемы, таблицы, диаграммы, что помогает интерпретации результатов анализа. Кроме того, пользователи могут анализировать базовую форму слов, учитывать грамматические особенности языка, так как приложение поддерживает функции морфологического анализа и лемматизации.

LancsBox можно использовать для создания корпуса исторических текстов, для исследования исторических изменений в языке путем анализа изменений в лексике, грамматике и других аспектов языка. Возможности LancsBox обсуждались в [2].

Результаты анализа диссертаций с применением приложения LancsBox

1. С помощью **частотного анализа** выявлены наиболее используемые в диссертациях слова. Например, слова *текст, коммуникация, язык, дискурс, анализ* встречаются с наибольшей частотой, что указывает на центральное место обозначаемых ими понятий в филологических исследованиях (см. табл. 1).

2. Проанализируем контекст употребления ключевого слова *дискурс* с помощью **конкордансного анализа** (см. табл. 2). Из приведенного примера видно, что термин *дискурс* употребляется в разных контекстах, приобретая различные определители: *устный, научный, политический, историографический* и т. д. Таким образом, прослеживается многомерность и широкое употребление термина в лингвистических исследованиях.

Таблица 1

Частотный анализ текстов диссертаций с помощью сервиса LancsBox

Слово	Число вхождений
текст	920
коммуникация	845
язык	812
дискурс	780
анализ	610
исследование	590
прагматика	560
семантика	530
конструкция	480
система	460
взаимодействие	430
когнитивный	408

Таблица 2

Конкордансный анализ текстов диссертаций с помощью сервиса LancsBox

Контекст до слова	Ключевое слово	Контекст после слова
…в данном контексте научный	дискурс	проявляется в анализе…
…анализируется развитие интернет-	дискурс	как новая форма коммуникации…
…в рамках исследования был изучен	дискурс	политический на примере речей…
…проведенный лингвистический анализ	дискурс	выявил признаки устного…
…межличностной коммуникации важен	дискурс	как форма выражения…
…в исследовании рассматривается историографический	дискурс	и его особенности в…
…анализ культурного	дискурс	выявил новые аспекты взаимодействия…

3. Приложение LancsBox дает возможность исследования **коллокаций**, что помогает выявить и проанализировать устойчивые словосочетания и фразеологизмы. Например, коллокация слова *семантика* представлена в таблице 3.

Таблица 3

Коллокационный анализ текстов диссертаций с помощью сервиса LancsBox

Коллокация	MIscore	t-score	Frequency
семантика слова	5.8	20.5	45
семантика предложения	5.5	18.3	38
семантическое поле	6.2	22.1	50
анализ семантики	5.9	19.8	42
семантика текста	5.3	17.5	35
семантика термина	5.0	16.2	32
синтаксическая семантика	4.8	15.4	30

Анализ показывает, что словосочетания *семантика слова, семантика предложения, семантическое поле* и *анализ семантики* являются устойчивыми в диссертациях.

4. Основные тематические направления диссертаций помогает выявить процедура определения **ключевых слов** и частоты их использования. Например, для темы «Прагматика» выделяются следующие ключевые слова и конструкции: *речевая ситуация, коммуникативный акт, прагматические маркеры* и т. д.

5. После проведенной процедуры генерации частотного списка выделим **семантические поля** (см. табл. 4).

Таблица 4

Примеры семантических полей, построенных с помощью сервиса LancsBox

Семантическое поле	Лексемы поля
Изучение текстов	текст анализ дискурс исследование прагматика семантика конструкция
Коммуникации	коммуникация взаимодействие дискурс прагматика
Языковые системы	язык система когнитивный семантика структура

Проанализируем семантическое поле лексемы *коммуникация* (см. табл. 5).

Таблица 5

Семантическое поле лексемы «коммуникация», построенное с помощью сервиса LancsBox

Ключевое слово	Контекст
коммуникация…	развитие средств массовой информации значительно изменило подходы к коммуникации…
коммуникация…	…исследование особенностей межкультурной и межличностной коммуникации выявило…
коммуникация…	в отличие от письменной коммуникации, устная коммуникация требует…

Ключевое слово — *коммуникация* — используется в различных контекстах: *межличностная коммуникация, массовая коммуникация, устная коммуникация*. В одном контексте часто появляются вместе слова из семантического поля «Коммуникация», что говорит о функциональных и тематических связях.

Генерация и анализ частотного списка с помощью программы LancsBox помогают выявить ключевые слова и связанные с ними семантические поля, что предоставляет ценную информацию для дальнейшего исследования и интерпретации текстов.

Анализ приложения для обработки текстов AntConc

AntConc — приложение для анализа и обработки текстов — разработано Лоуренсом Энтони, профессором прикладной лингвистики в Университете Васэда в Японии. Первая версия программы была выпущена в 2002 г. и с тех пор постоянно обновлялась, улучшалась и адаптировалась к нуждам пользователей.

Целью создания AntConc было предоставить исследователям и преподавателям лингвистики мощный бесплатный инструмент для анализа текстовых корпусов. Программа разработана таким образом, чтобы быть доступной и интуитивно понятной даже для пользователей без технического образования.

AntConc предлагает широкий спектр возможностей для анализа текстов. Основные функции включают: извлечение ключевых слов и анализ часто встречающихся слов и выражений в тексте, создание конкордансов, формирование списка всех встречающихся в тексте слов с их контекстами, сравнительный анализ корпусов, т. е. сравнение текстов друг с другом для выявления различий и сходств, подсчет частот встречаемости слов и выражений, исследование последовательностей слов в тексте; сегментация текста AntConc является кросс-платформенным приложением и может работать на различных операционных системах, включая Windows, macOS и Linux. Это делает его доступным для широкой аудитории пользователей. Программа поддерживает различные текстовые форматы, включая обычный текст, HTML, XML, а также форматы, используемые в корпусной лингвистике. Это позволяет пользователям работать с различными типами текстов без необходимости их предварительного преобразования [3; 4].

Одной из основных задач, выполняемых AntConc, является поддержка образовательных процессов. Программа активно используется в преподавании лингвистики, помогая студентам и исследователям анализировать текстовые данные и делать выводы на основе полученных результатов [5]. Кроме того, AntConc широко используется в научных исследованиях для анализа текстовых корпусов, изучения языковых закономерностей, исследования стилей письменной речи, анализа литературных текстов и др. Программа активно используется в корпусной лингвистике для анализа больших объемов текстов, что позволяет исследователям выявлять частотные и контекстные характеристики слов и выражений, а также проводить сравнительный анализ различных текстовых корпусов.

Одним из наиболее распространенных применений AntConc является анализ частотности слов в текстах. Например, исследователь может использовать программу для анализа частотности слов в корпусе научных статей, чтобы выявить наиболее часто используемые термины и выражения. Помимо этого, с помощью AntConc исследователи могут извлекать наиболее частотные ключевые слова, что может быть полезно для создания словарных статей, словарей, анализа тематического содержания и др. С помощью функции создания конкордансов можно определять все случаи употребления слова в контексте.

Результаты анализа диссертаций с применением инструмента AntConc

По итогам проведенного анализа был получен ряд результатов. Приложение AntConc продемонстрировало свою эффективность как инструмент корпусного анализа.

1. С целью определения ключевых тем и понятий в корпусе анализируемых диссертаций проведен **частотный анализ**. Анализ показал следующие частотно встречающиеся слова и словосочетания: *структура, язык, анализ, функция, текст, линвистическая теория, семантическое поле, языковая единица* и т. д. Это помогает понять основные методологические подходы и концепты, используемые в диссертациях.

2. С помощью **конкордансного анализа** было выявлено, как используются термины. Например, в текстах, описывающих процессы функционирования языка в обществе и социальные коммуникации, часто встречается термин *дискурс*. Пример результатов конкордансного анализа ключевого слова *дискурс* приведен в таблице 6.

Таблица 6

Конкордансный анализ лексемы «дискурс» с помощью приложения AntConc

Коллокация	Контекст
дискурс	…В данной диссертации изучается дискурс интернет-коммуникации в разносистемных языках…
дискурс	…Антропоцентрический дискурс является важной составляющей…
дискурс	…теоретические аспекты дискурса рассматриваются в главе…
дискурс	…влияние социальных сетей на письменный дискурс…
дискурс	…прагматика и дискурс тесно связаны в данном исследовании…
дискурс	…анализировал политический дискурс на предмет…
дискурс	…исторический дискурс и его эволюция были подробно рассмотрены…

Можно заключить, что термин *дискурс* употребляется в контекстах, связанных с влиянием социальных сетей, антропоцентризмом, интернет-коммуникацией. Частое использование рядом со словом атрибутивов *интернет-, письменный, исторический, политический* может указывать на различные подвиды дискурсов, анализируемых в диссертациях. Обнаруживается тесная связь между дискурсом и прагматикой, что отражает значимость прагматического анализа в филологических исследованиях.

3. **Коллокации** позволяют понять, какие понятия чаще всего ассоциируются друг с другом в научной литературе.

Предположим, ключевое слово для анализа — *дискурс* (см. табл. 7).

Таблица 7

Коллокационный анализ текстов диссертаций с помощью приложения AntConc

Ключевое слово	Коллокации	Число вхождений
дискурс	интернет-	54
дискурс	научный	37
дискурс	политический	29
дискурс	письменный	23
дискурс	антропоцентрический	18

Коллокации показывают наиболее часто встречающиеся сочетания слов с ключевым термином *дискурс*. Это помогает понять, в каких контекстах и значениях используется данный термин в диссертациях.

4. **Семантические поля:** частотные списки помогли выделить крупные семантические поля, использующиеся в диссертациях, — например поля, связанные с изучением текстов, коммуникаций и языковых систем.

Анализ приложения для обработки текстов WordSmith Tools

Приложение WordSmith Tools было разработано в начале 1990-х гг. Майком Скоттом, профессором лингвистики из Ливерпульского университета. Первоначальная цель создания приложения заключалась в предоставлении исследователям и филологам мощного инструмента для анализа текстов [6]. Программа была разработана с учетом потребностей академической среды, способствовала проведению лингвистических исследований и анализу больших объемов текстов.

Основной целью создания WordSmith Tools является автоматизация анализа текстов. Программа позволяет пользователям анализировать большие текстовые массивы без необходимости ручной обработки, что значительно экономит время и усилия. Еще одной важной целью создания приложения является предоставление возможности проведения углубленного лингвистического анализа текстов. WordSmith Tools позволяет исследовать структуру текста, частотность слов и фраз, а также их контекстное использование [7].

Программа поддерживает различные текстовые форматы, что делает ее универсальным инструментом для пользователей с разными потребностями. Одной из ключевых способностей приложения является анализ частотности слов и фраз в тексте. Это позволяет исследователям определять, какие слова и выражения наиболее часто используются в определенном тексте или корпусе текстов.

Исследователь может загрузить в программу текст произведений, например Шекспира, и определить, какие слова и фразы встречаются чаще всего. Это может помочь в изучении стиля автора и выявлении ключевых тематических элементов. Для группировки слов и фраз по отдельным признакам используется кластерный анализ, также данное приложение обладает функцией проведения статистического анализа текстов.

WordSmith Tools позволяет создавать и управлять корпусами текстов, что особенно полезно для исследователей, работающих с большими объемами данных, и позволяет генерировать детализированные отчеты по результатам анализа.

Результаты анализа диссертаций с применением инструмента WordSmith Tools

1. С помощью приложения были выделены самые **частотные** функциональные **слова**: *и, в, с, на, для, что, как*, а также часто встречающиеся термины *речь, язык, дискурс.*

2. При работе с **коллокациями** часто обнаруживаются фиксированные формулы, такие как *диалогический дискурс, прагматические функции, корпусный анализ*. Коллокации вокруг основного термина могут

свидетельствовать об основных исследовательских направлениях и аспектах.

Рассмотрим, какие коллокации возникают в корпусе диссертации по теме «Лингвистическая специфика интернет-дискурса в разносистемных языках» и какие мысли и направления исследований можно выявить на основе этих коллокаций. Так, были получены следующие коллокации: *интернет-дискурс + разговорный стиль, интернет-дискурс + формальный и неформальный, интернет-дискурс + жанры интернет-коммуникации, лингвистическая специфика + яркие признаки, лингвистическая специфика + грамматическая структура, лингвистическая специфика + лексические особенности, коммуникативные стратегии + в интернет-дискурсе, коммуникативные цели + пользователей, коммуникативная функция + интернет-языка* и т. д.

3. **Структурный анализ**: тексты диссертаций имеют четкую структуру, где введение и методология содержат более высокую концентрацию специализированных терминов, а обсуждение результатов показывает более разнообразные и менее формальные лексические единицы. Чтобы более детально проиллюстрировать структурный анализ, можно рассмотреть, как различные разделы диссертации концентрируют определенные типы лексических единиц и специализированных терминов. Например, во введении ставятся задачи исследования, определяются цели, раскрывается актуальность темы, формулируются гипотезы, поэтому часто встречаются специфичные для данной области термины, например *актуальность исследования, цель работы, задачи исследования, гипотеза, интернет-дискурс, лингвистическая специфика* и т. д.

Описание методов, используемых для проведения исследования, содержит значительное количество терминов, связанных с методикой анализа данных, статистическими методами и лингвистическими техниками. Например, *методы корпусного анализа, контекстуальный анализ, лингвистическая прагматика, лексические коллокации, частотные характеристики* и т. д.

В основной части исследования и в обсуждении результатов обычно содержатся более разнообразные и менее формальные лексические единицы, а также обобщения и объяснения. Например, *интересно отметить, варьируется, открывает новые перспективы, культурный контекст, проанализировав* и т. д.

Использование WordSmith Tools для анализа диссертаций по филологии предоставляет глубокое понимание лексических паттернов, методологических тенденций и общих направлений исследований, что способствует развитию и обогащению научного дискурса в этой области.

Анализ онлайн-платформы для обработки текстов «Текстометр»

«Текстометр» — это программное обеспечение, предназначенное для анализа текстов, позволяющее проводить семантический анализ, оценивать частотность слов и фраз, выявлять ключевые термины и др. [8; 9]. Одним из главных преимуществ «Текстометра» является высокая точность анализа текстов.

Результаты анализа диссертаций с применением инструмента «Текстометр» представлены в таблице 8.

Таблица 8

Анализ текстов диссертаций с помощью сервиса «Текстометр»

99 баллов из 100 (9.9)		
Структурная сложность	10 из 10	
Лексическая сложность	10 из 10	
Динамичность текста	0 из 10	
Описательность текста	10 из 10	
Знаков с пробелами	2 898 653	
Предложений	14 864	
Слов	337 052	
Уникальных слов	103 523	
Средняя длина слова	8.3	
Средняя длина предложения	27.9	
Лексическая плотность (8 из 10)		
Лексическое разнообразие (0.4)		
Частотный словарь по тексту	вопрос	380
	ответ	342
	диалогический	129
	дискурсивный	114
	темпоральный	87
	когнитивный	68
	язык	67
	модель	64
	художественный	63
	маркер	53
	единство	52
	исследование	48

Заключение (Conclusion)

В ходе проведенного исследования нами был осуществлен комплексный анализ основных программных инструментов для обработки текстов: AntConc, LancsBox, WordSmith Tools и «Текстометр». Особое внимание уделялось их применимости для анализа корпуса диссертаций, поскольку это требует специфических функций и высокой точности, необходимых для академических исследований.

AntConc выделяется своей простотой и доступностью. Он предоставил действенные средства для базового анализа текстов, такие как частотные словари, конкордансы и ключевые слова. AntConc удобен для быстрого анализа больших текстовых корпусов, несмотря на ограниченное количество дополнительных функций.

LancsBox показал себя как мощный и многофункциональный инструмент, подходящий для глубокого анализа текстов. Его гибкость и возможности визуализации данных оказались особенно полезными при исследовании сложных лингвистических структур в диссертациях.

WordSmith Tools отличается богатым набором средств для анализа текста, позволяющих детализировать и синтезировать результаты исследования. Он идеален для пользователей, занимающихся частотным анализом и стилометрией, хотя его коммерческая лицензия может стать ограничением для некоторых академических учреждений.

«Текстометр» способен работать в русскоязычной среде и предоставляет мощные инструменты для текстового анализа. Этот инструмент подходит для пользователей, работающих с русскоязычными диссертациями, предлагая специализированные функции для работы с этим языком.

Результаты нашего анализа могут служить ориентиром для исследователей и академических учреждений при выборе программного обеспечения для обработки текстов в зависимости от их задач и нужд. В конечном итоге правильный инструмент поможет повысить эффективность и надежность проводимых лингвистических исследований.

1. LancsBox: приложение : [сайт]. URL: https://lancsbox.lancs.ac.uk (дата обращения: 25.07.2024).

2. Позднякова Е. М., Суворина Е. В. Применение системы LancsBox в когнитивно-корпусных исследованиях // Когнитивные исследования языка. 2020. № 3 (50). С. 539–542.

3. Руководство по использованию приложения AntConc. 2018. Электрон. версия. URL: https://antconc-manual.readthedocs.io/_/downloads/en/latest/pdf/ (дата обращения: 25.07.2024).

4. [Корпусный анализ с AntConc] // Programming Historian : [сайт]. URL: https://clck.ru/3M93jW (дата обращения: 25.07.2024).

5. Котюрова И. А. Корпусные исследования с помощью сервиса AntConc в условиях работы в вузе // Язык и культура. 2020. № 52. С. 36–50. DOI: 10.17223/19996195/52/3

6. Interview with Mike Scott, WordSmith Tools developer // EFL Notes : [site]. 2016. April 18. URL: https://eflnotes.wordpress.com/2016/04/18/interview-with-mike-scott-wordsmith-tools-developer/ (дата обращения: 25.07.2024).

7. WordSmith Tools Manual version 3.0. 1998. The electronic version. URL: https://lexically.net/wordsmith/version3/manual.pdf (дата обращения: 25.07.2024).

8. Текстометр : онлайн-приложение. URL: https://textometr.ru/ (дата обращения: 25.07.2024).

9. Лапошина А. Н., Лебедева М. Ю. Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. 2021. Т. 19, № 3. С. 331–345. DOI: 10.22363/2618-8163-2021-19-3-331-345