Litera
Правильная ссылка на статью:
Падерина Т.С. Структуры данных для хранения языкового материала: принципы и оптимизация // Litera. 2025. № 11. DOI: 10.25136/2409-8698.2025.11.76630 EDN: FTZFMB URL: https://nbpublish.com/library_read_article.php?id=76630

Структуры данных для хранения языкового материала: принципы и оптимизация

Падерина Татьяна Сергеевна

ORCID: 0000-0002-2603-6242

младший научный сотрудник; лаборатория лингво-педагогических исследований; Федеральный исследовательский центр «Иркутский институт химии им. А.Е. Фаворского Сибирского отделения Российской академии наук»

664033, Россия, Иркутская обл., г. Иркутск, Свердловский р-н, ул. Лермонтова, д. 134, офис 120

jana-pad@mail.ru

Статья из рубрики "Автоматическая обработка языка"

DOI:

10.25136/2409-8698.2025.11.76630

EDN:

FTZFMB

Дата направления статьи в редакцию:

03-11-2025

Дата публикации:

10-11-2025

Аннотация: В данной статье представлен универсальный подход к созданию специализированного корпуса аннотированных данных, предназначенного для обучения модели извлечения информации из научной литературы по узкой специализации. Процесс включает в себя сбор данных, разработку принципов аннотирования с учетом лингвистических особенностей академических текстов и контекстуальных характеристик. Обсуждаются как общие структуры данных в программировании (массивы, списки, деревья), так и специализированные (корпусы, лексиконы, онтологии), адаптированные для решения лингвистических задач. Особое внимание уделяется принципам оптимизации, включая унификацию метаданных, многоуровневую разметку, обеспечение репрезентативности и поддержку мультимодальности. Теоретической базой послужили

работы по корпусной лингвистике, лингвосемиотическим основаниям изучения научного дискурса, проектированию лингвистических онтологий и разработке структур данных для лингвистических исследований. Рассматривается методология автоматической обработки текстов как неотъемлемый компонент работы с такими структурами, включая классификацию методов (статистические, на основе правил, машинное обучение) и этапы анализа (морфологический, синтаксический, семантический). Теоретическая значимость и практическая ценность работы заключается в том, что она вносит вклад в развитие корпусной лингвистики, в части изучения возможностей и проблем, возникающих в процессе корпусных исследований. Отмечается, что способность структур данных справляться с лингвистической неоднозначностью и отражать сложные взаимосвязи между языковыми элементами, используя механизмы логического вывода и принципы синергетики, является важным критерием для создания интеллектуальных систем. Предполагается дальнейшее использование обученной модели для автоматического извлечения данных из большого массива неразмеченной литературы, которые сформируют граф знаний предметной области. Подобный граф знаний открывает возможности для решения прикладных задач, включая составление частотных словарей по узким научным специальностям, отслеживание тенденции смены терминологического аппарата, в том числе появление новой терминологии.

Ключевые слова:

корпус текстов, языковой материал, научные данные, извлечение информации, автоматическая обработка текстов, принципы разработки, структура данных, разметка, аннотирование, граф знаний

Введение

Быстрый рост научно-технической информации детерминирует необходимость более тщательного освоения методов машинного обучения для решения задач эффективного сбора, систематизации, хранения, обработки и дальнейшего использования неструктурированного языкового материла по заданным параметрам для определённого массива текстов (корпуса) в автоматическом режиме. Разработка структур данных (= систем, прим. автора) для лингвистических исследований является основополагающим этапом, требующим детальной аннотации, стандартизации представления данных и поддержки анализа на всех языковых уровнях — от морфологии до прагматики [8].

Предметом нашего исследования является разработка системы хранения данных для обучения модели извлечения информации из подготовленного корпуса узкоспециализированных англоязычных научных текстов по направлению «Науки о Земле». Критерии отбора тестов определяется задачами Государственного задания, которое выполняет лаборатория и обеспечивают отбор релевантного контента, наиболее соответствующего требованиям письменной научной коммуникации [3,7].

Для эффективного использования, организации и обработки информации часто приходится работы с готовыми структурными данными, а иногда и разрабатывать их с нуля. В программировании структура данных определяется как способ организации информации для более эффективного использования, представляющий собой набор данных, связанных определенным образом. Главное свойство структур данных заключается в том, что каждая единица данных должна иметь четкое место, по которому ее можно найти, а характеристики структур включают определенное представление

данных в памяти и наличие алгоритмов для взаимодействия с ними, включая добавление и извлечение элементов [2].

Актуальность исследования принципов разработки структур данных для лингвистического (языкового) материала обусловлена необходимостью повышения качества и эффективности методов обработки текстов, таких как информационный поиск, фильтрация, рубрикация, кластеризация документов, автоматическое аннотирование и машинный перевод. Данная работа ставит целью систематизировать и обобщить принципы создания эффективных структур данных для языкового материала и рассмотреть методологические основы их автоматической обработки.

Понятие языкового материала и его особенности

Лингвистика XIX века ставила своей целью изучение языка как такового, в то время как лингвистика XXI века смещает акцент не столько на теоретическое знание, сколько на его прикладной аспект и практическое применение [13]. Проблема трудоемкого поиска научного материала потребовала, в свою очередь, оптимизировать работу в том числе с языковым материалом (тексты, слова, морфемы, фонемы, графемы) и как следствие, к появлению нового лингвистического направления – корпусная лингвистика. Основной объект современной корпусной лингвистики — текст — в различных формах своей реализации является одной из главных составляющих системы языка и речемыслительной деятельности.

В связи с многообразием научных текстов (в том числе на разных языках), разработка структур языковых данных требует тщательного отбора принципов и механизмов с учетом особенностей языкового материала, к которым относим следующие аспекты:

•***Многоуровневая структура:*** Язык может быть проанализирован на различных уровнях, таких как фонетико-фонологический, грамматический (морфологический и синтаксический), семантический и прагматический.

•***Неоднозначность и вариативность*:** Слова и выражения могут иметь несколько значений (полисемия, омонимия), а также различные формы употребления (словоформы, диалекты, социолекты).

•***Контекстуальная зависимость*:** Значение и использование языковых единиц часто зависят от контекста.

•***Динамичность*:** Язык постоянно меняется, что требует возможности актуализации и расширения данных.

•***Избыточность*:** Естественные коммуникационные системы обладают избыточностью для повышения надежности передачи информации и снижения недоразумений.

Таким образом мы рассматриваем языковой материал как совокупность дискретных взаимосвязанных и взаимообусловленных элементов, где определенное место отводится тексту как элементу для решения задач сбора, систематизации, обработки и дальнейшего использования неструктурированного языкового материла.

Специализированные структуры данных для лингвистического материала

Структуры данных являются фундаментальным понятием в программировании, позволяющим упорядочивать, искать, анализировать и использовать данные с применением алгоритмов. Они представляют собой способ организации информации для

более эффективного использования, где каждая единица имеет четкое место, по которому эту информацию можно найти по автоматическому запросу. Основные общие структуры данных представлены на рисунке 1:

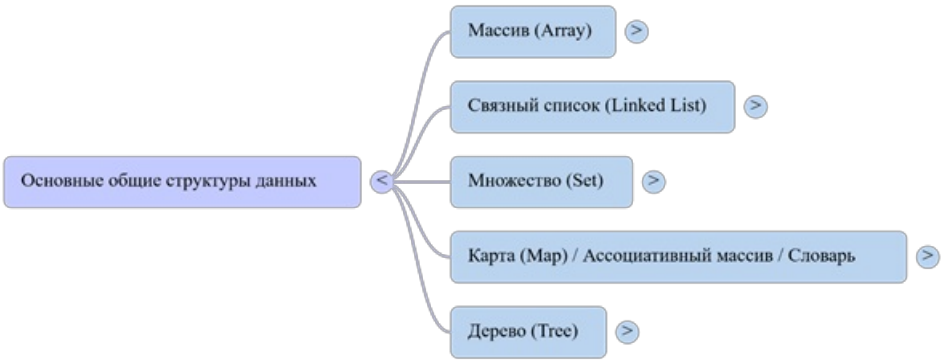

Рис.1 Общие структуры данных

Массив (Array), например, является простой структурой данных, на которой основаны многие другие, такие как списки, стеки и очереди. Каждый элемент массива имеет индекс, по которому его можно извлечь, а данные можно просматривать, сортировать и изменять. В свою очередь связный список (Linked List) представляет собой группу узлов, каждый из которых содержит данные и указатель на следующий узел (иногда на предыдущий). Позволяет быстро перемещаться между элементами, которые хранятся отдельно. Множество (Set) – схожи с классическими математическими множествами, используются для поддержания уникальных элементов, хранения несортированных данных и выполнения операций сравнения и объединения. Порядок элементов не имеет значения. В структуре Карта (Map) / Ассоциативный массив / Словарь данные хранятся в паре «ключ/значение», где каждый ключ уникален, а значения могут повторяться, что позволяет быстро искать данные по ключу. Частным случаем является хэш-таблица (hash-map), использующая хэш-функцию для вычисления индекса по ключу. Дерево (Tree), например, двоичное дерево поиска (Binary Search Tree) используется для быстрого поиска, а также для хранения данных в отсортированном виде с возможностью быстрого добавления и удаления. Префиксное дерево (так же Trie, prefix tree, бор) хранит данные последовательно, где каждый узел — это префикс, по которому находятся следующие узлы, что полезно для автозаполнения.

Для хранения и обработки сложного языкового материала используются более специализированные структуры, часто построенные на основе общих принципов организации данных, но адаптированные под лингвистические задачи:

1. Корпусы текстов. Корпус текстов — это подобранная и обработанная по определенным правилам совокупность текстов, используемых в качестве базы для исследования языка [9]. В современной лингвистике под корпусом обычно подразумевается корпус текстов в электронном виде, поскольку применение компьютерных технологий превратило его в эффективный исследовательский инструмент. Одно из основных свойств корпуса – репрезентативность. Корпус должен релевантно отображать объект, который моделирует (например, язык, диалект, жанр), путем сбалансированного включения различных типов текстов. Однако отмечаем, что добиться полной репрезентативности корпуса — труднодостижимая задача.

2. Лексиконы и Словари. В отличие от словаря, который обычно используется для обозначения списка всех слов, существующих в языке, лексикон, как понятие теоретического языкознания, используется для обозначения части лингвистической

теоретической модели данного языка. В этом смысле он тесно связана с «грамматической» (или синтаксической) частью модели, то есть с набором синтаксических, морфологических и фонологических правил, сформулированных теорией [17, с. 42–43].

3. Онтологии и Тезаурусы. За последнее десятилетие понятие *онтология* стало довольно распространённым в лингвистике. Вслед за рядом зарубежных и отечественных ученых будем рассматривать *онтологию* как «взаимосвязанную сеть релевантных концепций, которая раскрывает, классифицирует и упорядочивает допущения и термины рассматриваемой области» [4,5,6,14,15,19,21]. Объектом изучения в данном случае является не сама реальность, а человеческое восприятие. *Тезаурус,* в свою очередь — это нормативный словарь, который устанавливает лексические единицы и фиксирует семантические отношения между этими единицами в качестве элементов [11,12,16].

Основными целями тезаурусов является перевод естественного языка документов и запросов на контролируемый словарь, обеспечение последовательного использования единиц индексирования и описание отношений между терминами. Семантические сети являются основой для представления знаний в тезаурусах и онтологиях, отображая объекты и их связи.

4. Нейронные сети. Нейронные сети широко используются для обработки больших объемов экспериментальных и экспертных знаний. Они могут применяться для моделирования отклика системы на внешнее воздействие, классификации внутренних состояний, прогнозирования динамики, оценки полноты описания и оптимизации параметров. Обучение нейронных сетей может происходить «с учителем» (когда известно желаемое значение выхода) или «без учителя» (на основе наблюдения за входными значениями). В контексте лингвистики нейронные сети используются, например, для текстовой аннотации (для виртуальных помощников и чат-ботов), позволяя ИИ давать разумные ответы и поддерживать диалог, а также для распознавания объектов и эмоциональной окраски текста [3].

Методология создания корпуса данных для предметной области

В связи с затруднительным поиском готовых корпусов для узкоспециализированных научных областей, а зачастую с их отсутствием, разработка специализированных программ для выборки из массива информации конкретных данных, по запросу пользователя, на наш взгляд, нуждается в описании. В нашей работе мы предприняли попытку описать результат сбора и предварительной обработки англоязычных научных текстов по направлению подготовки «Науки о Земле» для дальнейшего использования, обработанного материла при обучении языковой модели. Эффективным методом, на наш взгляд, является разработка автоматизированной программы (разработка веб-сканер, от англ. web crawler, также встречается название «веб-паук»), которая позволит собрать данные с сайтов профильных журналов. Данный шаг позволяет нам собрать названия, аннотации, данные об авторах и другую информацию из статей, опубликованных в открытом доступе, а также выбрать для анализа публикации за определенный период времени.

Для тестовой проверки системы автоматического сбора данных, мы собрали более 50 статей, опубликованных в журнале Landslides издательства Springer по состоянию на 20 октября 2025 года. Landslides — международный журнал для публикации комплексных исследований, посвящённых оползневым процессам, анализу рисков, смягчению последствий и защите окружающей среды от последствий данной стихии. Данный этап

позволяет сформировать первоначальный массив текстов, из которого будет произведена выборка для дальнейшей разметки и аннотирования.

Принципы разработки и оптимизации структур данных для языкового материала

Для эффективной разработки структуры данных для лингвистического материала нами был предварительно проведен анализ теоретических положений, направленный на сопоставление и сравнение существующих систем хранения данных и принципов разработки, направленных на оптимизацию хранения, поиска и анализа. На основании проведенного анализа, и с учетом задач Государственного задания, нами были отобраны следующие ключевые, на наш взгляд, признаки.

1 . Неоднозначность языкового материала. Языковой материал характеризуется неоднозначностью, контекстуальной зависимостью значений и иерархической структурой:

· **Корпус**

· **Документ (статья, книга, отчет)**

· **Раздел/Глава**

· **Параграф**

· **Предложение**

· **Слово/Токен**

· **Морфема** (необязательно)

При разработке структур данных необходимо учитывать многоуровневые структуры (анализ на фонетическом, грамматическом, семантическом и прагматическом уровнях), вариативность (полисемия, омонимия, разнообразие форм употребления), контекст, что, в свою очередь предусматривает механизмы для снятия семантической омонимии, разрешения анафоры и кореферентности, а также фиксирования информационной структуры текста.

2 . Разметка и аннотирование. Разметка (англ. tagging, annotation) — процесс приписывания текстам и их компонентам специальных меток. Она дает возможность идентифицировать тексты по различным параметрам, позволяя осуществлять осмысленный поиск [10]. Высококачественная аннотация данных является одной из основных задач для точных моделей машинного обучения, пример разметке на рисунке 2.

Рис.2. Лингвистическая разметка (аннотирования) текста

Следует отметить, что к одному и тому же текстовому фрагменту может быть «привязано» множество разных видов аннотаций (Таблица 1), которые можно разделить на собственно лингвистические и экстралингвистические [1,2,20]:

Собственно лингвистические				Экстралингвисти (метаразметка)
Базовые аннотации (морфологические)	Синтаксические аннотации	Семантические аннотации	Дискурсивные аннотации	Метаданны
Токенизация: разделение текста на слова и знаки препинания.	Синтаксический разбор: определение грамматических связей между словами в предложении.	Именованные сущности (распознавание именованных сущностей, NER): выделение имён людей, названий организаций, географических объектов, дат, числовых значений.	Кореференция: связывание местоимений и других выражений с объектами, на которые они ссылаются.	Автор, названи публикации, ис жанр, язык, а технические (кодировка, обработки)
Лемматизация: приведение слов к словарной форме (лемме).	Деревья зависимостей: графическое представление синтаксических связей.	Семантические роли: определение ролей участников действия (агент, пациент, инструмент).	Структура дискурса: Разметка риторических отношений между частями текста.	Для устных ко пол, возраст, говорящего, записи
Частеречная разметка (POS-тегирование): определение части		Разрешение неоднозначности слова (Word Sense		

речи для каждого слова (существительное, глагол, прилагательное и т. д.).		Disambiguation, WSD): определение конкретного значения многозначного слова в данном контексте.		

Табл.1. Виды аннотаций

3 . Репрезентативность и сбалансированность. Корпус должен хорошо «представлять» моделируемый объект, что достигается путем специальной процедуры отбора текстов. Репрезентативность связана со сбалансированностью, что достигается путем специальной процедуры отбора текстов, сбалансированного включения различных типов текстов, а также их структурирования по жанрам, временным периодам и социолингвистическим характеристикам. Однако, считается, что полная репрезентативность в корпусах труднодостижимая задача.

4 . Гибкость и адаптивность. Гибкость в работе с различными языками и текстовыми структурами имеет решающее значение, особенно в условиях глобального исследовательского аспекта, когда многоязычная обработка данных становится необходимостью. Системы должны не только поддерживать различные языки, но и адаптироваться к различным диалектам и специализированным терминологиям, используемым в различных научных дисциплинах [3, с. 152]. Открытый характер стандарта (например, TEI, основанный на XML) позволяет доработку, расширение и адаптацию под новые задачи. Модульный подход в разработке систем обеспечивает независимую и распределенную реализацию, повторное использование и легкость расширения. Онтологии, в свою очередь, должны быть способны к обновлению и пополнению знаний.

5. Стандартизация, точность.

Для стандартизации и точности аннотирования текста с использованием XML в Python можно использовать библиотеку xml. etree., которая позволяет создавать и манипулировать XML-структурами. Ниже представлен код, который демонстрирует, как можно создать XML-документ для аннотирования текста с четкими значениями для каждого поля. Учитывая, что сам код достаточно объемный, в статье в качестве примера непосредственно использования мы приведем только часть кода:

```
import xml.etree.ElementTree as ET

class Annotation: <…..>

# Пример использования

annotation = Annotation()

# Добавление токенов (пример)

annotation.add_token("Уникальные", "ADJ", "amod", "структуры")

annotation.add_token("геологические", "ADJ", "amod", "структуры")

annotation.add_token("и", "CCONJ", "cc", "структуры")
```

```
annotation.add_token("гидрогеологические", "ADJ", "amod", "структуры")

annotation.add_token("структуры", "NOUN", "nsubj", "сделали")

annotation.add_token("сделали", "VERB", "ROOT","")

annotation.add_token("оползень", "NOUN", "nsubj", "сделали")

annotation.add_token("крайне", "ADV", "advmod", "неустойчивым")

annotation.add_token("неустойчивым", "ADJ", "acomp", "оползень")

# Установка подлежащего и сказуемого

annotation.set_subject("структуры")

annotation.set_predicate("сделали")

# Добавление сущностей (пример)

annotation.add_entity("ползень", "EVENT")

# Генерация XML

xml_output = annotation.to_xml ()

print (xml_output)
```

6. Параллельные корпуса: Для машинного перевода и сравнительной лингвистики часто используются параллельные корпуса, где один и тот же текст представлен на нескольких языках. Структуры данных должны поддерживать выравнивание (alignment) между предложениями, фразами или даже словами в разных языках. Это может быть реализовано через ссылки или общие идентификаторы. Пример:

```
{

"sentence_id": "s1",

"text_en": "These unique geological and hydrogeological structures have rendered the landslide mass highly unstable",

"text_ru": " Уникальные геологические и гидрогеологические структуры сделали оползень крайне неустойчивым",

"alignment": [

{"en_ru": [0, 2], "ru_en": [0, 6]}, // these -> Уникальные

{"en_ru": [4, 8], "ru_en": [7, 12]} // unique -> геологические

// ... и так далее

]

}
```

7 . Модульность и микросервисная архитектура. Одной из ключевых особенностей современных информационных систем является применение архитектуры, основанной на микросервисах. Такая архитектура разделяет функциональные компоненты системы — импорт, предварительную обработку, аннотирование, структурирование и визуализацию данных – на независимые модули, что повышает гибкость, надежность и масштабируемость системы. Каждый модуль может быть разработан, обновлен и протестирован отдельно, что позволяет значительно ускорить процесс интеграции новых алгоритмов обработки и поддержки дополнительных типов медиа. Такая модульная организация также способствует удобству управления доступом, так как разные уровни пользователей могут иметь разграниченные права на доступ и внесение изменений в базу данных.

8. Поддержка контроля версий и регулярное обновление. Одним из ключевых факторов успешной реализации структур данных является обеспечение регулярного обновления и контроля версий хранимого материала. Применение специализированных протоколов для управления изменениями, а также периодическая проверка корректности всех метаданных позволяют поддерживать высококачественный и актуальный корпус. Такой процесс необходим для обеспечения непрерывности научных исследований и образовательных программ, поскольку данные постоянно обновляются за счет поступления новых материалов и корректировки уже существующих.

9 . Валидация данных и обеспечение семантической целостности. Для того чтобы структурированные данные могли быть использованы как для академических исследований, так и для прикладных задач, необходимо обеспечить их валидацию и семантическую согласованность. Это достигается за счёт регулярной проверки корректности аннотаций, согласованности морфологических и синтаксических меток, а также контроля за правильностью заполнения метаданных. В идеале, системы должны поддерживать функции автоматической сверки данных с эталонными наборами и проводить экспертный анализ выявленных расхождений, что является важным аспектом в разработке надёжных лингвистических корпусов.

10. Масштабируемость и гибкость доступа. Особое внимание в современных структурах данных уделяется возможности расширения и масштабирования системы по мере увеличения объёма данных. Это требует от разработчиков обеспечения гибкости хранилища, возможности быстрого индексирования, адаптации к увеличению нагрузки и оперативного обновления функционала. Масштабируемость достигается за счёт применения распределённых систем хранения, облачных технологий и алгоритмов оптимизации запросов, что позволяет эффективно работать с большими массивами языковых данных в условиях постоянно растущего спроса на вычислительные ресурсы [3, с. 151–153].

Заключение

Таким образом, принципы разработки структур данных для хранения языкового материала охватывают широкий спектр технических, методологических и организационных аспектов. Разработка структур данных для хранения языкового материала является междисциплинарной задачей, объединяющей лингвистику, компьютерные науки и искусственный интеллект [18]. Благодаря междисциплинарному синтезу теоретических знаний и современных технических решений возможно создание таких систем, которые не только сохранят огромное разнообразие языковой информации, но и станут основой для дальнейших исследований и практического применения в области перевода, языкового моделирования и автоматизированного анализа речи

В рамках нашего исследования, на основании отобранных в автоматическом режиме и предварительно размеченных научных текстов, определен материал для обучения языковой модели для текстов узкой специализации. В дальнейшем мы предполагаем, что обученная модель позволить нам формировать графы знаний предметных областей в автоматическом режиме и решать прикладные задачи (создавать частотные словари, работать с терминологическим аппаратом, находить новые паттерны и генерировать связные тексты по заданному научному направлению).

Библиография

1. Падерина Т. С. Автоматическое извлечение ключевых терминов из корпуса научных статей в SCP // Верхневолжский филологический вестник. 2024. № 3(38). С. 139-144. doi: 10.20323/2499-9679-2024-3-38-139. EDN: PUBMIE.
2. Костюшкина Г. М., Свердлова Н. А., Баребина Н. С. и др. Лингвосемиотические основания изучения научного дискурса. Москва: Общество с ограниченной ответственностью "ФЛИНТА", 2024. 216 с. ISBN 978-5-9765–5690-4. EDN: TPHYAF.
3. Падерина Т. С. Методы извлечения терминов в научных текстах (на материале статей по направлению науки о земле) // Казанский лингвистический журнал. 2023. Т. 6, № 3. С. 388-396. DOI 10.26907/2658–3321.2023.6.3.388-396. EDN: VLCYAH.
4. Бурков А. А. Инженерия машинного обучения / пер. с англ. А. А. Слинкина. Москва: ДМК Пресс, 2022. 306 с.
5. Чилингарян К. П. Корпусная лингвистика: теория vs методология // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2021. Т. 12. № 1. С. 196-218. doi: 10.22363/2313-2299-2021-12-1-196-218. EDN: YMIAME.
6. Риз Р. Обработка естественного языка на Java / Р. Риз; пер. с англ. А. В. Снастина. 2-е изд., эл. Москва: ДМК Пресс, 2023. 266 с.
7. Junger J. Predicate formation in the verbal system of Modern Hebrew. Amsterdam: University of Amsterdam, 1987. 183 p.
8. Кравцов Д. В., Коростелев Д. А., Юркова О. Н. Автоматизированная система для построения онтологий предметных областей // Мониторинг. Наука и технологии. 2017. № 1(30). С. 46-50. EDN: YNCCNP.
9. Лукашевич Н. В., Добров Б. В. Проектирование лингвистических онтологий для информационных систем в широких предметных областях // Онтология проектирования. 2015. Т. 5, № 1(15). С. 47-69. EDN: TOPTMZ.
10. Наместников А. М., Пирогова Н. Д., Филиппов А. А. Подход к автоматическому построению лингвистической онтологии для определения интересов пользователей социальных сетей // Онтология проектирования. 2021. Т. 11, № 3(41). С. 351-363. DOI 10.18287/2223–9537-2021-11-3-351-363. EDN: JVKREP.
11. Fabry P., et al. Rethinking Meaning and Ontologies from the Perspective of Ontological Units. 27 Mar. 2025.
12. Gendron B., et al. Towards Ontology-Based Descriptions of Conversations with Qualitatively-Defined Concepts. 1, arXiv, 2025. https://doi.org/10.48550/ARXIV.2509.04926.
13. Schalley A. C., Musgrave S., Haugh M. Accessing phonetic variation in spoken language corpora through non-standard orthography // Australian Journal of Linguistics. 2014. 34(1), 139-170. https://doi.org/10.1080/07268602.2014.87545.
14. Zhang D., et al. Meronymic Ontology Extraction via Large Language Models. 1, arXiv, 2025. https://doi.org/10.48550/ARXIV.2510.13839.
15. Федотов А. М., Идрисова И. А., Самбетбаева М. А., Федотова О. А. Использование тезауруса в научно-образовательной информационной системе // Вестник

Новосибирского государственного университета. Серия: Информационные технологии. 2015. Т. 13, вып. 2. С. 86-102. EDN: UJEXAX.
16. Федюченко Л. Г. Терминологическая база данных как трансферная модель технического знания: специальность 10.02.21 "Прикладная и математическая лингвистика": диссертация на соискание ученой степени доктора филологических наук. 2021. 407 с. EDN: PUSZLE.
17. ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri. 2nd ed. Geneva: International Organization for Standardization, 1986.
18. Соловьева А. Е. Англоязычные тексты военной авиации как основа лингвистического корпуса // Балтийский гуманитарный журнал. 2019. Т. 8, № 3(28). С. 369-372. DOI 10.26140/bgz3-2019–0803-0093. EDN: MSLXAE.
19. Антопольский А. Б. Международная стандартизация в сфере управления лингвистическими информационными ресурсами // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2021. № 5. С. 23-32. DOI 10.36535/0548–0027-2021-05-5. EDN: ORIEWZ.
20. Zeroual I., Lakhouaja A. Data Science in Light of Natural Language Processing: An Overview // Procedia Computer Science. 2018. Vol. 127. Pp. 82-91. Crossref, https://doi.org/10.1016/j.procs.2018.01.101.
21. Lust B., Blume M., Pareja-Lora A., Chiarcos C. Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences: An Introduction. Cambridge: The MIT Press, 2020. https://doi.org/10.7551/mitpress/10990.003.0002.

Результаты процедуры рецензирования статьи

Рецензия выполнена специалистами Национального Института Научного Рецензирования по заказу ООО "НБ-Медиа".
В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов можно ознакомиться здесь.

В статье последовательно раскрывается предмет исследования, который строго соответствует теме. Исследование посвящено анализу современных лингвистических методов обработки информации, особенностям аннотирования и структурирования языковых данных. Все основные разделы статьи соотносятся с её заглавием, обеспечивая логическую согласованность между темой и содержанием работы.
Используется широкий спектр современных методов лингвистических исследований, приведены примеры обработки лингвистических данных, элементы XML-аннотирования, алгоритмы, что подтверждает высокий уровень методологической проработки. Методические подходы описаны подробно, что положительно сказывается на прозрачности и воспроизводимости научных результатов.
Актуальность темы статьи очевидна: развитие цифровой лингвистики, автоматизации языкового анализа в XXI веке обеспечивает значимость вопросов, рассматриваемых в статье, для лингвистической науки и практики.
В статье отчётливо прослеживается научная новизна: предложен сравнительный обзор алгоритмов и систем аннотирования, освещены новые подходы к обработке многослойной языковой информации, приведён анализ актуальных отечественных и зарубежных работ по теме. Особое внимание уделено теоретическим вопросам, связанным с многоуровневым анализом текстов, что расширяет современные представления в области цифровой лингвистики.
Статья отличается чёткой структурой: введение, постановка целей, основной

аналитический материал, выводы. Содержание изложено строгим научным стилем, терминология используется корректно. В тексте присутствуют таблицы, схемы, коды программ, что облегчает восприятие сложного материала. Логика изложения выдержана, переходы между частями статьи обоснованы и последовательны.

Библиографический список содержит более 20 источников: представлены современные зарубежные публикации, работы на русском языке, большие корпусные исследования, а также программные стандарты и документы. Все источники процитированы в тексте статьи.

В заключительной части подводятся итоги исследования, указываются перспективы дальнейшей работы, делаются выводы о необходимости развития цифровых лингвистических ресурсов. Материал будет интересен, прежде всего, специалистам в области цифровой лингвистики и студентам профильных направлений. Для широкой научной аудитории статья может послужить источником информации о современных тенденциях развития лингвистических исследований.

Комментарии по языку статьи:

В тексте не выявлено существенных орфографических ошибок. Отмечается достаточно высокая стилистическая культура изложения, что характерно для работ академического уровня. Однако присутствуют отдельные погрешности пунктуационного характера (например, непоследовательное выделение вводных слов: таким образом, в свою очередь; в предложении, начинающемся со слов: «Однако, считается, что полная репрезентативность…» - слово «однако» вводным не является, поэтому в запятой после него нет необходимости). Замечены небольшие грамматические неточности и опечатки («Множество (Set) – схожи с классическими математическими множествами, используются для…», «Основными целями тезаурусов является…», «модель позволить нам формировать», «приходится работы» (вместо «приходится работать»), а также слово «тестов» вместо «текстов» во втором абзаце статьи.

Статья может быть рекомендована к публикации после небольшой стилистической правки, что повысит её привлекательность в глазах читателей академического круга.