

2020-12-16

Исследовательский прототип параллельного корпуса научно-технических текстов

Представлен исследовательский прототип параллельного корпуса научно-технических текстов. Описано устройство разрабатываемого параллельного корпуса. Показаны роли пользователей в параллельном корпусе. Рассмотрены виды разметки, отражающие особенности научно-технических текстов. Особое внимание уделено терминологической разметке. Предложены источники научно-технических текстов для наполнения корпуса.

Разрешение многозначности поискового запроса в корпусе научно-технических текстов

Одним из факторов, влияющих на релевантность результатов информационного поиска, является многозначность поискового запроса, выраженная лексическими средствами естественного языка. Отмечено, что практически каждое слово естественного языка, кроме строго однозначных терминов, особенно принадлежащее общеупотребительной лексике, многозначно, его зависит от контекста употребления слова. Многозначность лексической единицы проявляется на этапе поискового запроса. Предлагается подход к разрешению многозначности лексических единиц в поисковом запросе, что позволит решить проблему лексической многозначности и существенно разгрузить поисковую выдачу, оставив только предметную область, представляющую интерес для пользователя.

Темпоральный аспект гендерной вариативности английского произношения (корпусное исследование)

Статья посвящена темпоральному исследованию гендерной вариативности английского произношения. Показано, что гендерные исследования в лингвистике занимают важное место, что подтверждается количеством и качеством проведенных исследований. Обосновано, что гендерные исследования по фонетике требуют использования специального программного обеспечения, а возможность пользователя создавать собственные речевые подкорпуса не только значительно сокращает время на сбор и обработку иллюстративного материала, но и расширяет спектр проводимых исследований. Описана разработанная система разметки звучащей речи, представлены ее основные компоненты, а также исследовательские возможности созданного на ее основе речевого корпуса. В качестве примера проведено темпоральное исследование речи юношей и девушек из городов Белфаст, Лондон, Брэдфорд, Лидс и Кембридж, Дублин, Ньюкасл и Кардифф. Выявлены различия между речью юношей и девушек при чтении текста, а также показаны региональные различия на темпоральном уровне.

Анализ современных корпусов параллельных текстов

В последние годы создание корпуса и тематические исследования стали неотъемлемой частью деятельности лингвистов. Методология корпуса становится частью лингвистической науки, и лингвисты, работающие в различных областях, как правило, проводят свои исследования на базе корпусов. Данная статья содержит информацию о параллельных корпусах, их разновидностях, об основных аспектах создания и использования параллельного корпуса.

Особенности разметки в речевом корпусе для распознавания диалектов русского языка

Статья посвящена особенностям разметки в речевом корпусе для распознавания диалектов русского языка. Показано, что интерес к созданию речевых корпусов был в значительной степени инициирован разработками в области автоматического распознавания речи, где исследователям приходится сталкиваться с огромной акустической вариативностью звуковых единиц языка, что подтверждается количеством и качеством проведенных исследований. Обосновано, что для распознавания диалектов русского языка требуют использования специального программного обеспечения по созданию речевых корпусов. Описана разработанная система разметки звучащей речи, представляющая собой веб-приложение, которое позволяет пользователю с помощью браузера взаимодействовать с речевой базой

данных. Показано, что задача система разметки речи учитывает характерные черты каждого диктора, влияющие на качество произношения. Показано, что дикторы из одного региона примерно одного возраста произносили одни и те же слова по-разному.

2021

Модель учебно-научного текста для разметки корпуса научно-технических текстов

В статье предложена модель структуры учебно-научного текста для разметки корпуса научно-технических текстов. Обоснована необходимость учитывать композиционную структуру научно-технических текстов при разметке корпуса. Отмечено, что учебно-научные тексты имеют одинаковую для всех текстов этого класса структуру изложения материала, а также содержат ограниченный набор структурных элементов. Охарактеризованы структурные элементы учебно-научного текста. Представлена композиционная структура учебно-научных текстов в нотациях Бекуса-Наура. Предложена модель учебно-научного текста в виде графа, вершинами и ребрами которого являются полноценные структурные элементы учебно-научного текста.

2022-04-28

Базовый шаблон многоязычной словарной статьи предметной онтологии на основе корпуса научно-технических текстов

Рассмотрены аспекты разработки шаблона многоязычной словарной статьи предметной онтологии. В частности, представлен шаблон многоязычной словарной статьи лингвистической онтологии предметной области на основе корпуса научно-технических текстов для русского и английского языков. Шаблон может быть использован для решения задач автоматической обработки естественного языка. Базовый шаблон многоязычной словарной статьи лингвистической онтологии в предметной области на основе научно-технических корпусов быть полезна для решения проблем, связанных с задачами обработки естественного языка научно-технических текстов.

Особенности перевода заголовков с русского на английский язык (на материале научно-технических текстов по космонавтике)

Рассмотрены особенности перевода заголовков с русского на английский язык на материале заголовков научно-технических текстов по космонавтике. Проанализированы способы формулирования русскоязычных заголовков. Заголовки проклассифицированы по общим моделям, выделены общие характеристики научно-технических заголовков, продемонстрированные на примерах. Изучены структурные и лексические особенности перевода заголовков научно-технических статей с русского на английский язык.

2022-12-16

Методика наполнения словарной статьи терминологической базы данных на основе параллельного корпуса научно-технических текстов

Рассмотрены процессы создания лингвистической онтологии предметной области и ранжирования лингвистических словарных статей онтологии. Разработан и представлен базовый шаблон словарной статьи предметной онтологии для русского и английского языков на основе научно-технических текстов. Приведен анализ научно-технического термина по данному шаблону. Шаблон может быть использован для решения задач автоматической обработки естественного языка.

Моделирование семантико-синтаксической структуры предложений для разметки корпуса научно-технических текстов

Рассмотрена актуальная на сегодняшний день проблема по аспектам моделирования семантико-синтаксической структуры предложений для разметки корпуса научно-технических текстов. Рассмотрены основные подходы по делению семантических падежей. За основу взята падежная грамматика Ч. Филлмора. Описаны валентность и семантические падежи на примере научно-технических текстов. Проанализированы семантические падежи в текстах по космонавтике и авиации. Описаны результаты исследования и представлены в таблице.

Особенности семантической разметки в корпусе научно-технических текстов

Рассматривается проблема разметки семантических отношений между компонентными предложениями в корпусе научно-технических текстов. Проанализированы основные подходы, описывающие предметно-предикатные отношения в предложении. Обосновано, что наибольшую сложность при семантической разметке корпуса научно-технических текстов вызывают многокомпонентные термины. Описан подход к выделению семантической разметки научно-технических текстов на основе семантических падежей Ч. Филлмора.

Терминологическая разметка научно-технических текстов в специальном корпусе

Представлен результат лингвистического эксперимента, цель которого сравнить данные ручной разметки и данные, полученные в ходе автоматической разметки фрагмента научного текста. Эксперимент проводили с помощью таких онлайн ресурсов, как TermExtraction, TerMine и Terminology Extraction. Проведенное исследование показало, что результаты, полученные в ходе автоматической разметки, не являются точными, по сравнению с разметкой текста, проведенной вручную. Однако использование онлайн ресурсов позволяет значительно сократить время при работе с объемным материалом.

2023-04-18

Особенности перевода терминов-эпонимов с русского на английский язык в параллельных текстах

При переводе эпонимов с русского языка на английский имена ученых часто изменяются. Для того чтобы рассмотреть такие случаи, решено проанализировать английские и русские эпонимы в предметной области «Космонавтика» и изучить способы передачи русских эпонимов на английский язык. Для этого выбраны эпонимы из параллельных текстов и проведено сравнение этих терминов. По результатам найденных терминов было обнаружено, что фамилии в эпонимах были изменены или включали иностранные символы того или иного языка, так как сохраняли культуру стран ученых.

2023-04-21

Роль мультидисциплинарных исследований в автоматической обработке научно-технических текстов

Работа посвящена реализации мультидисциплинарного подхода для решения задач автоматической обработки научно-технических текстов в части извлечения специальной терминологии, который был опробован в ходе реализации практик для студентов кафедр Л4 и ИУ7. Описаны задачи и результаты полученные на каждом этапе, а также поделены аспекты исследования, реализуемые студентами разных кафедр. Показаны результаты анализа эффективности разработанной системы извлечения многокомпонентных терминов из параллельных текстов. Представлены междисциплинарные трудности реализации такого подхода, а также некоторые перспективы его развития.