Министерство образования и науки

Российской Федерации

ФГБОУ ВО «Тверской государственный университет»

Факультет Иностранных языков и Международной коммуникации

Направление «Лингвистика»

Кафедра английского языка

РЕФЕРАТ

На тему: «Одноязычные корпусы»

|  |  |
| --- | --- |
|  | Выполнила:  Студентка группы 14, дневная форма обучения  Яранцева Юлия Сергеевна |
|  |  |
|  | Проверила: кфн, доцент, доцент кафедры английского языка  Масленникова Е.М. |

ТВЕРЬ

2021

Содержание

|  |  |
| --- | --- |
| Определение понятия «корпус» ………………………………… | 3 |
| Отличия корпусной лингвистики от традиционной лингвистики | 3 |
| Типы лингвистических корпусов …………………………………. | 4 |
| Определение понятия «одноязычный корпус» ………………...… | 4 |
| Одноязычные корпуса, которые могут быть полезны русскоговорящему пользователю …………………………………. | 5 |

**Определение понятия «корпус»**

Известно, что корпусная лингвистика, как раздел компьютерной лингвистики, который занимается разработкой общих принципов построения и использования лингвистических корпусов с использованием компьютерных технологий, появилась в 60-е годы XX века, преимущественно на основе английского языка, однако в это же время очень быстро начали возникать корпуса на основе других языков. Например, в Брауновском университете США в 1963 году ученым У.Н. Френсисом был создан первый корпус текстов на электронном носителе(<http://corpus.leeds.ac.uk/protected>), в котором содержалось 500 текстов 15 самых популярных жанров англоязычной прозы США по 2000 в каждом. К корпусам прилагались указатель частотности, а также некоторые статистические распределения.

Таким образом, получается, что под термином корпус подразумевается следующее: **корпус** – это собрание текстов одного или нескольких языков, которые подобраны и обработаны по определённым правилам текстов и которые используются в качестве базы для исследования языка. Корпус представляет собой собрание письменных и устных высказываний. Данные корпуса, как правило, хранятся на компьютерах и всевозможных электронных носителях, и потому они доступны в электронном виде. При этом составные части корпуса, тексты, состоят из данных, а также, в некоторых случаях, метаданных, которые описывают эти данные, и из лингвистических аннотация, которые эти данные упорядочивают.

**Отличия корпусной лингвистики от традиционной лингвистики**

Табл. 1. Отличия корпусной лингвистики от традиционной лингвистики (автор: В.В. Рыков)

|  |  |
| --- | --- |
| **Корпусная лингвистика** | **Традиционная лингвистика** |
| Основное внимание – изучение **речи** | Основное внимание – изучение **языка** |
| Цель – описание языка в том виде, как он проявил себя в речи, представленной в виде специально подобранного корпуса текстов | Цель – описание и объяснение языка |
| В своих исследованиях опирается на данные корпуса текста | В своих исследованиях идёт от теории к её объяснению и подтверждению в фактах речи |
| Предпочитает **квантитативные** (**количественные**) методы | Предпочитает **квалитативные** (**качественные**) методы |
| Видит себя частью традиций, базирующихся на **эмпирических** методах | Видит себя частью традиций, базирующихся на **рационалистических** методах |
| Текст рассматривается как некоторая **физическая сущность** | Текст рассматривается как некоторая **абстракция** |
| Составление грамматики **конкретных** языков | Изучает языковые **универсалии** |
| Основное внимание уделяется **форме** | Основное внимание – не только форме, но и **содержанию** |
| Рассматривает тексты в **глобальной** перспективе | Рассматривает тексты в **локальной** перспективе |
| Фокусирует своё внимание на как можно более **широком** взгляде на текст, неограниченном ни какими догмами | Анализирует некоторую **конкретную**, искусственно ограниченную, проблемную область |
| В своих выводах опирается на наблюдение речевой деятельности, проявленной в виде текстов | Опирается на интуицию в отборе речевого материала, в отборе эмпирических материалов своих исследований |
| Часто пользуется **вероятностными** **методами** и **статистикой** для первичной обработки речевого материала | Предпочитает **логические рассуждения** |
| Проводится работа с лингвистическими данными (словоупотреблениями) в том виде, в каком они встречались в контексте | Предпочитаются искусственные примеры, из изолированных от текста словоупотреблений |
| Предпочитает **индуктивные** методы обработки эмпирического словесного материала, считает их сутью научного метода | Предпочитает **дедуктивные** методы обработки эмпирического словесного материала |
| Верит в научные открытия, основанные на обработке эмпирических данных | Верит в открытия, основанные на процедурах, оценках, сравнениях и так далее |

**Типы лингвистических корпусов**

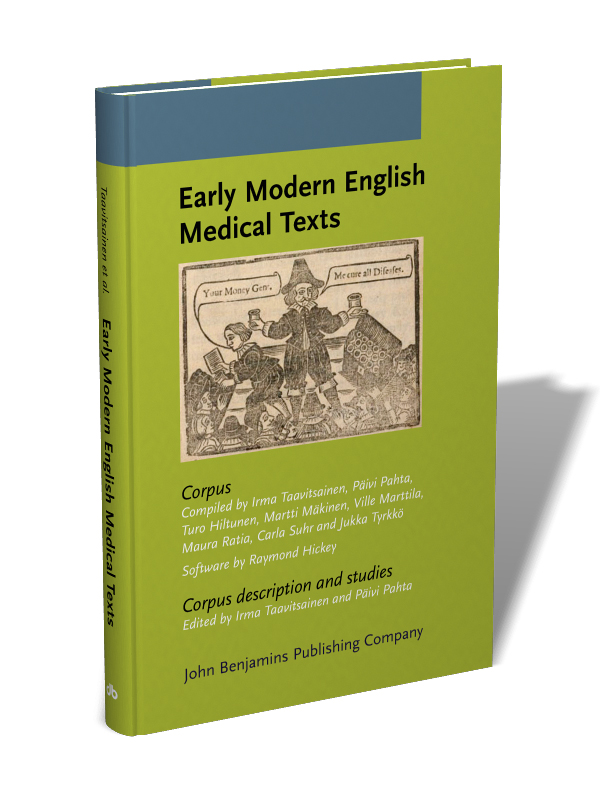
Табл. 2. Типы лингвистических корпусов (автор: Д. Панфилов)

|  |  |
| --- | --- |
| **Признак** | **Типы корпусов** |
| Тип языковых данных | Письменные, Устные, Смешанные |
| «Параллельность» | Одноязычные, Двуязычные, Многоязычные |
| «Литературность» | Литературные, Диалектные, Разговорные, Терминологические, Смешанные |
| Цель | Многоцелевые, Специализированные |
| Доступность | Свободно доступные, Коммерческие, Закрытые |
| Назначение | Исследовательские, Иллюстративные |
| Динамичность | Динамические (мониторные), Статические |
| Разметка | Размеченные, Неразмеченные |
| Характер разметки | Морфологические, Синтаксические, Семантические и другие |
| Объём текста | Полнотекстовые, «Фрагментнотекстовые» |

**Определение понятия «одноязычный корпус»**

В целом, классифицировать корпусы можно по различным признакам: цель создания корпуса, тип языковых данных, «литературность», жанр, динамичность, тип разметки, объём текстов и так далее. По критерию параллельности, например, корпусы можно разделить на одноязычные, двуязычные и многоязычные. Именно данные корпуса оказывают помощь переводчику в процессе редактирования и выделения трудностей в понимании текста, и воспроизведении прагматического потенциала оригинала при переводе.

Среди одноязычных корпусов можно выделить две группы: с одной стороны, корпуса, которые охватывают весь язык, с другой – те, которые охватывают только язык для специальных целей. Например, Corpus of Early English Medical Writing (CEEM) (Рис. 1) – это корпус медицинских текстов на английском языке 1375–1750 годов, общий объём которого насчитывает около 1,5 миллиона слов и в котором содержатся теоретические работы, справочники, стихотворные тексты на медицинские темы.

Помимо этого, в таком виде корпусов, как одноязычные, присутствуют различные варианты языка и диалекты. Английский носителей языка и английский лиц, говорящих на иностранном языке, не изучались до появления новейших компьютерных технологий, которые позволили привнести в лингвистическую науку контрастивный анализ большого количества сопоставляемых текстов. Основной функцией многих корпусов, таких как британский, американский или русский, является поиск сочетаний слов в предложении, то есть подбор оптимальных вариантов в языке перевода. Зачастую одноязычные корпусы языков предоставляют разъяснения и термины, которые могут быть незнакомы для целевой аудитории читателей (или для самого переводчика), а также готовые фрагменты из языка оригинала, которые могут помочь переводчику в выборе эквивалента в родном языке. Существует такая возможность, что и после анализа приводимых примеров специалист не может выполнить перевод термина, слова или фразы. В таком случае переводчик может использовать готовые отрывки с неизвестным элементом в качестве примера для анализа, в процессе которого на помощь приходят словари, энциклопедии или мнение экспертов в этой области. Подобные поисковые системы, позволяющие выделить примеры из культуры языка оригинала, имеют чрезвычайно важное значение.

**Одноязычные корпуса, которые могут быть полезны русскоговорящему пользователю**

Рис. 1. Corpus of Early English Medical Writing (CEEM)

1. Национальный корпус русского языка. (<https://ruscorpora.ru/>)

Национальный корпус русского языка (Рис. 2) – это корпус, созданный для научных исследований и обучения языку. Кроме того, данный корпус имеет две важные особенности: во-первых, он характеризуется представительностью, или сбалансированным составом текстов, что означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и что все эти тексты входят в корпус по возможности Изображение выглядит как текст

Автоматически созданное описаниепропорционально их доле в языке соответствующего периода.

Рис. 2. Страница поиска словоформ «Национального корпуса русского языка»

Однако, данный корпус предоставляет доступ не только к документам, в которых была использована та или иная словоформа, но и к статистике частотности употреблений искомой словоформы как в прозе, так и в поэзии.

1. Русский учебный корпус (<http://web-corpora.net/RLC>)

Данный корпус (Рис. 2) подходит под определение второй группы одноязычных корпусов (см. Определение понятия «одноязычный корпус»). Он содержит в себе образцы устной и письменной речи двух категорий нестандартных говорящих на русском языке: изучающих русский язык как иностранный и так называемых эритажных говорящих. Для первой категории русский язык не является родным, представители же второй категории начали усваивать его в детстве как первый язык, но по разным причинам в качестве основного языка общения используют другой язык.

Таким образом, корпус позволяет производить поиск по лексико-грамматическим свойствам, а также изучить такие сложные явления, как нестандартное использование русского вида, падежей, предложных конструкций, а также неверное использование слов или неверное понимание смысла в многословных конструкциях.

Рис. 3. Страница поиска в «Русском учебном корпусе»

1. Корпус русских учебный текстов (КРУТ) (<http://web-corpora.net/learner_corpus>)

Изображение выглядит как текст, снимок экрана, зеленый

Автоматически созданное описаниеКорпус русских учебных текстов (КРУТ) (Рис. 3) – это коллекция текстов на русском языке, написанных студентами разных вузов. Общий объем корпуса составляет около 3,1 миллиона слов. Тексты сопровождаются несколькими типами разметки (метатекстовой, морфологической разметкой и разметкой по ошибкам), что позволяет осуществлять поиск по корпусу.

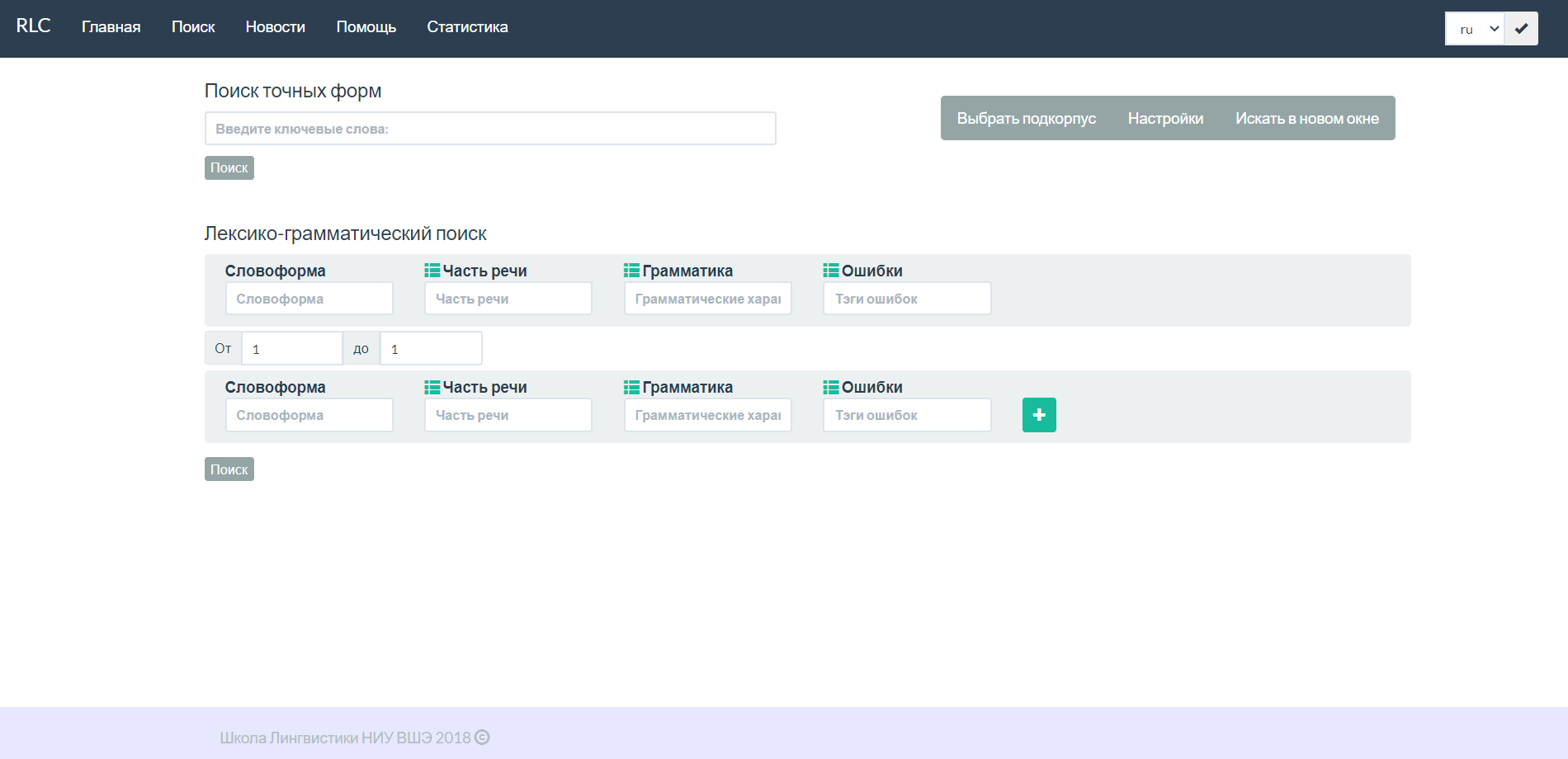
Корпус русских учебных текстов является информационно-справочной системой, предназначенной для исследователей, преподавателей, студентов, а также для всех, кто интересуется проблемами современной русской грамматики, актуальными процессами в области лексики, морфологии и синтаксиса современного русского языка.

Рис. 4. Страница поиска в «Корпусе русских учебных текстов»

Список использованной литературы:

Козлова Н.В., Лингвистические корпуса: определение основных понятий и типология, Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2013. Том 11, выпуск 1

Корпус русских учебный текстов (КРУТ) (<http://web-corpora.net/learner_corpus>)

Куренко К.Н., Корпусная лингвистика в переводоведении, Иностранные языки: лингвистические и методические аспекты Выпуск 37. 2017

Кутузов А.Б., Курс «Корпусная лингвистика», ТюмГУ

Национальный корпус русского языка (<https://ruscorpora.ru/>)

Панфилов Д., Корпусы текстов и принципы их создания, Спецсеминар «Искусственный интеллект» Кафедры АЯ ВМК МГУ

Русский учебный корпус (<http://web-corpora.net/RLC>)