

Министерство науки и высшего образования Российской Федерации  
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Филологический факультет

УТВЕРЖДЕНО:

Декан филологического факультета  
И.В. Тубалова

Оценочные материалы по дисциплине

**Технологии корпусной лингвистики**

по направлению подготовки

**45.03.03 Фундаментальная и прикладная лингвистика**

Направленность (профиль) подготовки:  
**«Фундаментальная и прикладная лингвистика»**

Форма обучения

**Очная**

Квалификация

**Бакалавр**

Год приема

**2024**

СОГЛАСОВАНО:

Руководитель ОП  
А.В. Васильева

Председатель УМК  
Ю.А. Тихомирова

Томск – 2024

## **1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами**

Целью дисциплины является изучение основных принципов и методов автоматической обработки текстов на естественном языке (ЕЯ)

Целью освоения дисциплины является формирование следующих компетенций:

ПК-4. Способен разрабатывать программный код при решении задач автоматической обработки текстов

ОПК-2. Способен к ведению профессиональной деятельности с опорой на основы математических дисциплин, необходимых для формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.1. Применяет способы формализации и алгоритмизации поставленных задач в сфере автоматической обработки текстов

ИПК-4.2. Создает программный код с использованием языков программирования и манипулирования данными в сфере автоматической обработки текстов

ИОПК-2.3 Применяет математические методы в процедурах анализа и синтеза лингвистических структур

## **2. Оценочные материалы текущего контроля и критерии оценивания**

К числу форм контроля, оценивающих уровень достижения компетенций по текущей дисциплине, относятся тесты, устные вопросы, выполнение практических заданий (ПК-4, ОПК-2, ИПК-4.1, ИПК-4.2, ИОПК-2.3), контрольная и самостоятельная работы (ПК-4, ОПК-2).

Итоговый проект включает блок вопросов по теоретической части изучаемой дисциплины (ПК-4, ОПК-2, ИПК-4.1, ИПК-4.2, ИОПК-2.3):

Пример практических задач:

Примерные задания Зарегистрируйтесь на сайте Национального корпуса русского языка по адресу: <https://rusc>

gr

ga.ru/ и выполните следующие задания:

1. Как выбрать нужный подкорпус или оформить поисковый запрос в зависимости от задач исследования?

- Изучаем концепт СЧАСТЬЕ в русской языковой картине мира. Наши действия?
- Хотим отследить частотность употребления репрезентантов концепта СЧАСТЬЕ

за последние два столетия.

- Социолингвистическое исследование – изучение специфики женской и мужской речи и особенности концептуализации. Концепт СЧАСТЬЕ.

- Репрезентация концепта СЧАСТЬЕ в прозе Виктории Токаревой.

- Особенности репрезентации концепта СЧАСТЬЕ в медиа картине мира за последние 10 лет. В художественной картине мира за последние 100 лет?

- Изучаем лингвокультурный типаж Бизнесмен в СМИ.

- Использование сокращений в текстах бытового (обиходного) дискурса.

- Использование существительных – диминутивов с отрицательной коннотацией в текстах обиходно-бытового дискурса (в блогах, в SMS-сообщениях).

- Выясняем особенности функционирования в медиадискурсе вводных оборотов со значением чувств говорящего (к счастью, к несчастью, к сожалению, странное дело, к удивлению)

- Изучаем семантику и функционирование местоимений-существительных в дательном падеже, обозначающих людей, в конструкции с предикативами (типа мне неловко, вам обидно, нам пора и т.д.) в устной речи.

- Изучаем метафору метеонима «дождь».

2. Выбрать два задания из предложенных и выполнить, опираясь на данные НКРЯ: • Какие идиомы с названиями частей тела (нога / рука / др. на выбор) используют носители современного русского языка? Какие из этих фразеологизмов используются чаще, чем другие?

- Как употребляются глаголы грузить – загрузить – загружать? Являются ли эти формы взаимозаменяемыми во всех контекстах или существуют такие употребления, где замена невозможна?

- В каких конструкциях используется наречие очевидно? Какие грамматические и дискурсивные функции оно выполняет в разных типах контекстов?

3. Разработайте собственный корпус, скачав массив текстов. Аргументируйте свой выбор корпуса. Придумайте гипотезу и решите, основываясь на методах количественной лингвистики, следующие задачи: частота лексем в корпусе, вхождение и сравнение лексемы или коллокатов. Результаты исследования отразите в презентации.

Ответы:

Теоретические вопросы (ПК-4, ОПК-2, ИПК-4.1):

Критерии оценивания ответа на теоретический вопрос

| Оценка       | Критерии                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | <ol style="list-style-type: none"> <li>1. Понимание и логика высказывания изученного материала</li> <li>2. Представление взаимосвязей процесса и взаимосвязи теоретических модулей изучаемого предмета</li> <li>3. Полнота данных ответов;</li> <li>4. Аргументированность данных ответов;</li> <li>5. Правильность ответов на вопросы;</li> </ol>                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| «зачтено»    | <p>Полно и аргументировано даны ответы по содержанию задания. Обнаружено понимание материала, может обосновать свои суждения, применить знания на практике, привести необходимые примеры не только по учебнику, но и самостоятельно составленные. Изложение материала последовательно и правильно</p> <p>Ответы обучающегося удовлетворяют тем же требованиям, что и для оценки «отлично», но допускается 1-2 ошибки, которые сам же исправляет.</p> <p>Обучающийся демонстрирует знание и понимание основных положений данного задания, но: 1) излагает материал неполно и допускает неточности в определении понятий или формулировке правил; 2) не умеет достаточно глубоко и доказательно обосновать свои суждения и привести свои примеры; 3) излагает материал непоследовательно и допускает ошибки.</p> |
| «не зачтено» | <p>Демонстрация незнания ответа на соответствующее задание, допускаются ошибки в формулировке определений и правил, искажающие их смысл, беспорядочно и неуверенно излагается материал; отмечаются такие недостатки в подготовке, которые являются серьезным препятствием к успешному овладению последующим материалом.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |

Критерии оценивания практической работы:

| Оценка       | Критерии                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | <ol style="list-style-type: none"> <li>1. Понимание и логика алгоритма работы</li> <li>2. Наличие или отсудив ошибок в коде</li> <li>3. Полнота решения практических задач</li> <li>4. Своевременность выполнения;</li> <li>5. Умения связать практический материал с теоретическим;</li> <li>6. Понимание базовых формул обработки естественного языка и программирования;</li> </ol>                                                                                                                                                                                                                                                                                                                                                                                                 |
| «зачтено»    | <p>Основные требования к решению практических задач выполнены. Продемонстрированы умение анализировать алгоритмы и находить оптимальное количество решений, умение работать с информацией, в том числе умение затребовать дополнительную информацию, необходимую для уточнения реализации алгоритма, навыки разработки программного кода;</p> <p>Основные требования к решению практических задач выполнены, но при этом допущены недочеты. В частности, недостаточно раскрыты навыки стиля, недостаточно комментариев</p> <p>Имеются существенные отступления от решения практических задач. В частности отсутствуют навык и умения моделировать решения в соответствии с заданием, представлять различные подходы к разработке алгоритмов, ориентированных на конечный результат</p> |
| «не зачтено» | Задача не решена, обнаруживается существенное непонимание проблемы                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |

Итоговый проект (ПК-4, ОПК-2, ИПК-4.1, ИПК-4.2, ИОПК-2.33)

Итоговый проект включает создание корпуса текстов. К числу тем могут относиться тексты социальных сетей, отзывы, литературных произведений и т.п..

Выбор тем текстов предусматривает возможность выбора индивидуальной образовательной траектории, связывающую проектную деятельность со смежными гуманитарными дисциплинами. Однако существуют ограничительные критерии реализации проекта, направленные на формирование вышеуказанных компетенций: 1. Сбор и структуризация данных. 2. Наличие двух классов в обучающем датасете, наличие числовых переменных.

Сбор и структуризация данных предусматривает свободу выбора темы. В случае, если выбор не был осуществлен, то предлагаются следующие рубрики:

1. Новостные сайты и пресс релизы (классификация и генерация текстов по рубрике и/или информационному агентству)
2. Отзывы к товарам или фильмам (Наличие 3 класса: позитивный, негативный, нейтральный отзывы)
3. Комментарии в социальных сетях
4. Стены сообществ «ВКонтакте» (классификация по рубриками)

Критерии оценивания практической работы:

| Оценка | Критерии                                                                                                                                                                                                               |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | <ol style="list-style-type: none"> <li>1. Сформулирована гипотеза проекта</li> <li>2. Собран и структурирован датасет</li> <li>3. Четкая логика реализации алгоритмов обработки естественного языка в коде;</li> </ol> |

|              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | <p>4. Правильность ответов на вопросы;</p> <p>5. Наличие структурированной презентации</p> <p>6. Полнота проекта</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| «зачтено»    | <p>Выполнены все требования к проекту: сформулирована гипотеза, создан прототип сайтов для сбора и структуризации информации, написан код для обработки естественного языка</p> <p>Выполнены все требования к составлению презентаций: дизайн слайдов, логика изложения материала, текст хорошо написан и сформированные идеи ясно изложены и структурированы</p> <p>Существуют незначительные ошибки в проекте, не влияющие на конечный результат. В частности, может быть низкий уровень формальных метрик, неточности в визуализации данных</p> <p>Основные требования к презентациям выполнены, но при этом допущены недочеты. В частности, имеются неточности в изложении материала; отсутствует логическая последовательность в суждениях; не выдержан объем презентации</p> <p>Существенные ошибки в коде, отсекает логика представления проекта, недостаточный объем датасета</p> <p>Имеются существенные отступления от требований к презентациям. В частности: тема освещена лишь частично; допущены фактические ошибки в содержании презентаций или при ответе на дополнительные вопросы.</p> |
| «не зачтено» | <p>Критические ошибки в коде, гипотеза не подвержена, недостаточное количество или отсутствие обучающей выборки</p> <p>Тема презентации не раскрыта, обнаруживается существенное непонимание проблемы</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |

### 3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

**Зачет в пятом семестре** состоит из трех частей.

**Первая часть** представляет собой тест из 20 вопросов, проверяющих ИПК-4, ИПК-4.2, ИОПК-2.3

Ответы на вопросы первой части даются путем выбора из списка предложенных, установления соответствия между объектами или формулирования однословного ответа на вопрос открытого типа.

#### Примерный тест

1. Лингвистический корпус. Принципы функционирования.

Примерные задания

1) Для корпусной лингвистики метаданные – это ... . Отметьте правильные ответы. балла

- жанр текста
- хронотоп текста
- морфологическая разметка

- сфера функционирования текста
- семантика
- данные об авторе
- синтаксис

2) Как называются программные средства, используемые для синтаксической разметки?

3) Разметка (аннотирование) – это ...

- обработка текста в целях облегчения его изучения.
- необходимый этап создания лингвистического корпуса.
- сведения, размещенные на главной странице корпуса.
- присвоение каждому элементу текста особого тэга.

4) В каком типе корпусов производится разметка ошибок?

- в эритажном корпусе
- в параллельном корпусе
- в мультимодальном корпусе
- в акцентологическом корпусе

5) Какие корпуса из перечисленных относятся к мегакорпусам?

- Aranea
- BNC
- Russian Web corpus
- RuTenTen
- НКРЯ
- ГИКРЯ
- Брауновский корпус

6) Утверждение «Мультимодальный корпус требует нескольких уровней разметки» предполагает, что ...

- элементы корпуса, представляющего визуальные образы, должны быть описаны иначе, чем элементы письменного или аудио текста.
- для корпуса МУРКО в составе НКРЯ были разработаны дополнительные

принципы

описания текстов, например, создана разметка значений жестов.

• мультимодальные корпуса должны быть частично закрытыми из соображений этики.

• разметка по цели высказывания или по дискурсивным элементам не нужна в обычных текстовых корпусах.

7) В корпусе с неснятой грамматической омонимией словоформа «уже» будет отнесена к ...

- одной лемме
- двум леммам
- трем леммам
- четырём леммам

8) На данных какого словаря основаны многие автоматические анализаторы морфологии русского языка?

9) Семантическая разметка в НКРЯ реализована в виде

- строгой онтологии
- нестрогой онтологии, совмещающей древовидный и фасеточный принципы организации

10) В качестве минимального элемента разметки в НКРЯ принята ..., то есть графическая единица, не обладающая, в отличие от леммы, семантическими характеристиками

11) Источником автоматической семантической разметки часто служат компьютерные ...

- энциклопедии
- данные
- тезаурусы
- программы
- приложения

12) Пропорциональное представление в корпусе текстов разных типов называется

...

13) Материалом мультимедийного корпуса служат

- аудиозаписи
- видеозаписи
- тексты
- все варианты

14) Неслучайные сочетания слов в корпусной лингвистике называются ....

15) Корпус текстов, созданных людьми, которые регулярно используют неродной язык для общения, называется ...

- корпусом ученических текстов
- корпусом эритажного языка
- корпусом лингва франка
- параллельным корпусом

16) Как Вы думаете, что означает утверждение «Корпус – это интерпретированный текст»?

- при помощи разметки создатели корпуса передают пользователю свои теоретические установки.
- прежде, чем включать тексты в корпус, лингвисты тщательно изучают их и делают выводы об их содержании.
- начиная работу в корпусе, пользователь по умолчанию соглашается принять на веру принципы, положенные в основу аннотации.
- каждое языковое явление должно быть интерпретировано создателями корпуса как правильное или неправильное и соответствующим образом помечено.

Критерии оценивания тестирования:

| Оценка       | Критерии                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|--------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | <ol style="list-style-type: none"> <li>1. Полнота выполнения тестовых заданий;</li> <li>2. Своевременность выполнения;</li> <li>3. Правильность ответов на вопросы;</li> <li>4. Самостоятельность тестирования</li> </ol>                                                                                                                                                                                                                                                                                                                                                                  |
| «зачтено»    | <p>Выполнено более 85 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос</p> <p>Выполнено более 70 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос; однако были допущены неточности в определении понятий, терминов и др.</p> <p>Выполнено более 54 % заданий предложенного теста, в заданиях открытого типа дан неполный ответ на поставленный вопрос, в ответе не присутствуют доказательные примеры, текст со стилистическими и орфографическими ошибками.</p> |
| «не зачтено» | Выполнено не более 53 % заданий предложенного теста, на поставленные вопросы ответ отсутствует или неполный,                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |

|  |                                                                              |
|--|------------------------------------------------------------------------------|
|  | допущены существенные ошибки в теоретическом материале (терминах, понятиях). |
|--|------------------------------------------------------------------------------|

#### **4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)**

Первая часть. Решение тестовых задач:

Тест №1

1. Основные теоретические положения и понятия корпусной лингвистики

Примерные задания

1) Корпусная лингвистика – это раздел ... . \*

- количественной лингвистики
- компьютерной лингвистики
- паралингвистики
- лингвистической статистики
- металингвистики

2) Сформулируйте объект корпусной лингвистики \*

3) Сформулируйте предмет корпусной лингвистики \*

4) Верно ли утверждение, что корпусный метод – это разновидность общенаучного метода наблюдения, который состоит в систематическом рассмотрении большого объема

естественных данных? \*

- Верно
- Не верно

5) Верно ли утверждение, что корпусный метод – это «пучок» методов из разных областей лингвистических исследований \*

- Верно
- Не верно

6) Кому из российских лингвистов принадлежит идея о том, что достоверные лингвистические данные могут быть получены лишь из большого массива текстов?

\*

- Владимиру Александровичу Плузгину
- Андрею Александровичу Кибрику
- Раймунду Генриховичу Пиотровскому
- Виктору Павловичу Захарову
- Анатолию Николаевичу Баранову

7) Какие утверждения справедливы по отношению к корпусной лингвистике? \*

• Цель – описание языка в том виде, как он проявил себя в речи, представленной в

виде

специально подобранного корпуса текстов

- Цель – описание и объяснение языка
- Предпочитает количественные методы
- Предпочитает качественные методы

9

• Опирается на интуицию в отборе речевого материала, в отборе эмпирических материалов своих исследований

• В своих выводах опирается на наблюдение речевой деятельности, проявленной в

виде

текстов

• Предпочитает дедуктивные методы обработки эмпирического словесного материала

- Предпочитает индуктивные методы обработки эмпирического словесного материала,

считает их сутью научного метода

8) Этот корпус с 1990-х годов считается образцовым \*

- COCA
- British National Corpus
- NOW
- Lexicum
- COSMAS
- НКРЯ
- Frantext

9) Отметьте свойства лингвистического корпуса (в узком значении) \*

- электронный вид
- коллекция текстов может быть не размечена
- репрезентативность
- обязательная разметка
- сбалансированность
- наличие системы поиска
- доступность
- обязательная возможность создания собственного подкорпуса

10) Коллигация – это \*

- регулярно воспроизводимые словосочетания со слабой степенью идиоматичности
- привычные выражения, состоящие из цепочек слов, которые часто используются вместе, трафаретные, шаблонные выражения
- совокупности морфолого-синтаксических условий, обеспечивающих

сочетаемость

языковых единиц

11) Что такое конкорданс? \*

12) Система управления текстовыми и лингвистическими данными - это \*

- KWIC
- корпус-менеджер
- база данных
- система поиска данных в корпусе

13) Выделите типы метаданных \*

- лингвистические
- экстралингвистические
- данные о структуре текста
- информационные
- лингвистические, описывающие элементы текста

14) Каковы этапы лингвистического анализа текста? \*

- лемматизация, токенизация, частичная и полная морфологическая разметка, синтаксическая разметка

10

- токенизация, лемматизация, частичная и полная морфологическая разметка, синтаксическая разметка
- токенизация, частичная и полная морфологическая разметка, синтаксическая разметка, лемматизация
- токенизация, лемматизация, синтаксическая разметка, частичная и полная морфологическая разметка

15) Специализированные корпуса – это \*

- корпуса, ограниченные один или несколькими жанрами
- двуязычные и многоязычные корпуса

- мультимодальные корпуса
- корпуса устной речи

16) Как называются корпуса, которые периодически обновляются, пополняются? \*

17) Верно ли утверждение, что все корпуса являются фрагментотекстовыми? \*

- Верно
- Не верно

18) В каком типе корпусов применяется метод автоматического выравнивания текста? \*

- специализированный
- мультимодальный
- просодический
- параллельный
- терминологический

## Практические задания

1. Блок заданий работы с корпусами

Зарегистрируйтесь на платформе English corpora по адресу:

<https://www.englishcorpora.org/corpora.asp> и выполните следующие задания:

1) Какой корпус на платформе выбрать в зависимости от задач исследования?

- Нам нужно изучить и сопоставить данные на материале текстов интернет-дискурса на английском языке США, Канады, Шри-Ланка и Австралии.
- Изучаем современный американский английский.
- Изучаем историю американского английского языка, эволюцию абсолютных конструкций.

• Анализируем средства речевого воздействия на материале речей, произнесенных в Британском Парламенте.

• Изучаем жанровые характеристики решений верховного суда.

• Исследуем устную речь, интересуют речевые акты угрозы в американском английском (британском английском).

• Изучаем образы мужчины и женщины в ранненовоанглийском языке.

• Исследуем новые дискурсивные практики, появившиеся с 2020 г.

2) Где посмотреть список текстов, их распределение по жанрам и объем в конкретном корпусе?

3) Как искать по лемме?

4) Как искать по словоформе?

5) Как просмотреть конкорданс слова? Контекст употребления слова?

6) Как искать по части речи?

7) Как найти синонимы слова?

8) Можно ли осуществлять поиск с помощью символов (подстановочных знаков)?

9) Выбрать два задания и выполнить, опираясь на данные корпуса (корпусов):

• Как функционирует слово tweet в современном английском языке? Сравните функционирование этого слова в BNC, COHA, COCA, NOW. Когда оно впервые появилось в современном значении? В каких контекстах оно употреблялось до появления Твиттера?

• Проанализируйте поведение словоформы sahoots. В каких типах текстов она используется? В каком синтаксическом окружении она встречается? Что можно сказать о ее семантической просодии?

• Как используется выражение snail mail? В текстах каких типов и жанров оно встречается? В каком окружении?

• Как функционирует выражение take ownership of? С какими существительными оно сочетается?

• Как функционирует глагол hamper? В каких формах он чаще всего используется? В каких синтаксических конструкциях? Как ведет себя существительное hamper?

- Является ли слово open в сочетании со словом door чаще глаголом или существительным?
- С какими существительными используется конструкция be bound to?
- Какие слова заканчиваются на -side, -phobia, -mania, -esque, -proof? В каких типах текстов они преимущественно встречаются?

Блок заданий разработки корпуса

Скачайте собственный корпус текстов в виде txt файлов. Загрузите в корпусный менеджер (AntConc, SketchEngine). Произведите анализ языковых явлений

Создайте корпус на базе движка Tsakorpus, предварительно разметив файлы в программе ELAN.

### **Информация о разработчиках**

Степаненко Андрей Александрович, старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики