Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Российский государственный университет им. А.Н. Косыгина
(Технологии. Дизайн. Искусство)»

Институт Магистратура
Кафедра общего и славянского искусствознания

# ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

для проведения текущей и промежуточной аттестации
по учебной дисциплине

## Корпусная лингвистика

| | |
|---|---|
| Уровень образования | магистратура |
| Направление подготовки | 44.04.01 Педагогическое образование |
| Программа | Дистанционные технологии в гуманитарном образовании |
| Срок освоения образовательной программы | 2 года 6 месяцев |
| Форма обучения | очно-заочная, заочная |

Оценочные материалы учебной дисциплины «Корпусная лингвистика» основной профессиональной образовательной программы высшего образования, рассмотрены и одобрены на заседании кафедры, протокол № 6 от 06.03.2023 г.

Составитель оценочных материалов учебной дисциплины:

профессор Г.В. Варакина

Заведующий кафедрой: Г.В. Варакина

## 1. ОБЩИЕ СВЕДЕНИЯ

Учебная дисциплина «Корпусная лингвистика» изучается на первом курсе
Курсовая работа/Курсовой проект – не предусмотрены.

Форма промежуточной аттестации:
зачет с оценкой

## 2. ЦЕЛИ И ЗАДАЧИ ОЦЕНОЧНЫХ СРЕДСТВ, ОБЛАСТЬ ПРИМЕНЕНИЯ

Оценочные средства являются частью рабочей программы учебной дисциплины и предназначены для контроля и оценки образовательных достижений обучающихся, освоивших компетенции, предусмотренные программой.
Целью оценочных средств является установление соответствия фактически достигнутых обучающимся результатов освоения дисциплины, планируемым результатам обучения по дисциплине, определение уровня освоения компетенций.
Для достижения поставленной цели решаются следующие задачи:

– оценка уровня освоения общепрофессиональных и профессиональных компетенций, предусмотренных рабочей программой учебной дисциплины;
– обеспечение текущего и промежуточного контроля успеваемости;
– оперативного и регулярного управления учебной, в том числе самостоятельной деятельностью обучающегося;
– Соответствие планируемых результатов обучения задачам будущей профессиональной деятельности через совершенствование традиционных и внедрение инновационных методов обучения в образовательный процесс.

Оценочные материалы по учебной дисциплине включают в себя:

– перечень формируемых компетенций, соотнесённых с планируемыми результатами обучения по учебной дисциплине*;*
– типовые контрольные задания и иные материалы, необходимые для оценки результатов обучения;

Оценочные материалы сформированы на основе ключевых принципов оценивания:

– валидности: объекты оценки соответствуют поставленным целям обучения;
– надежности: используются единообразные стандарты и критерии для оценивания достижений;
– объективности: разные обучающиеся имеют равные возможности для достижения успеха.

## 3. ФОРМИРУЕМЫЕ КОМПЕТЕНЦИИ, ИНДИКАТОРЫ ДОСТИЖЕНИЯ КОМПЕТЕНЦИЙ, СООТНЕСЁННЫЕ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕИСПОЛЬЗУЕМЫЕ ОЦЕНОЧНЫЕ СРЕДСТВА

| **Код компетенции, код индикатора достижения компетенции** | **Планируемые результаты обучения по дисциплине** | **Наименование оценочного средства** | |
|---|---|---|---|
| | | **текущий контроль (включая контроль самостоятельной работы обучающегося)** | **промежуточная аттестация** |
| УК-4<br>ИД-УК-4.1 | - редактирует различные академические тексты | **Письменная работа: Технология обработки текстового материала (технология критического мышления)**<br>**Круглый стол 1**<br>**Круглый стол 2**<br>**Устный опрос**<br>**Тестирование** | зачет с оценкой |
| УК-4<br>ИД-УК-4.2 | - демонстрирует готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации | **Письменная работа: Технология обработки текстового материала (технология критического мышления)**<br>**Круглый стол 1**<br>**Круглый стол 2**<br>**Устный опрос**<br>**Тестирование** | |

| **Код компетенции, код индикатора достижения компетенции** | **Планируемые результаты обучения по дисциплине** | **Наименование оценочного средства** | |
|---|---|---|---|
| | | **текущий контроль (включая контроль самостоятельной работы обучающегося)** | **промежуточная аттестация** |
| УК-4<br>ИД-УК-4.3 | - осуществляет межличностное деловое общение, в том числе на иностранных языках с применением профессиональных языковых форм и средств | **Круглый стол 1**<br>**Круглый стол 2**<br>**Устный опрос** | |

\

## 4. ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ И ДРУГИЕ МАТЕРИАЛЫ, НЕОБХОДИМЫЕ ДЛЯ ОЦЕНКИ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ И УРОВНЯ СФОРМИРОВАННОСТИ КОМПЕТЕНЦИЙ

**4.1.** Оценочные материалы **текущего контроля** успеваемости по учебной дисциплине, в том числе самостоятельной работы обучающегося, типовые задания

**УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.
**ИД-УК-4.1** Подготовка и редактирование различных академических текстов.
**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.
**ИД-УК-4.3** Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств.

**Письменная работа: Технология обработки текстового материала (технология критического мышления)** (УК-4, ИД-УК-4.1, ИД-УК-4.2)

**УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.
**ИД-УК-4.1** Подготовка и редактирование различных академических текстов.
**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.

Письменная работа
Время выполнения 90 мин
Форма работы – самостоятельная, индивидуальная.
А) Модель работы с концептуальной таблицей

1. Выделить основания для сопоставления (первый столбик).
2. Провести сравнительный анализ по выделенным основаниям.

| **Критерии для сопоставления объектов** | **Национальный корпус русского языка** | **Фундаментальная электронная библиотека (ФЭБ)** |
|---|---|---|
| **Общее** | | |
| 1. | | |
| 2… | | |
| **Различное** | | |
| 1. | | |
| 2. … | | |

Материал для сопоставления:
Национальный корпус русского языка и Фундаментальная электронная библиотека (ФЭБ).

**ОТВЕТ:**

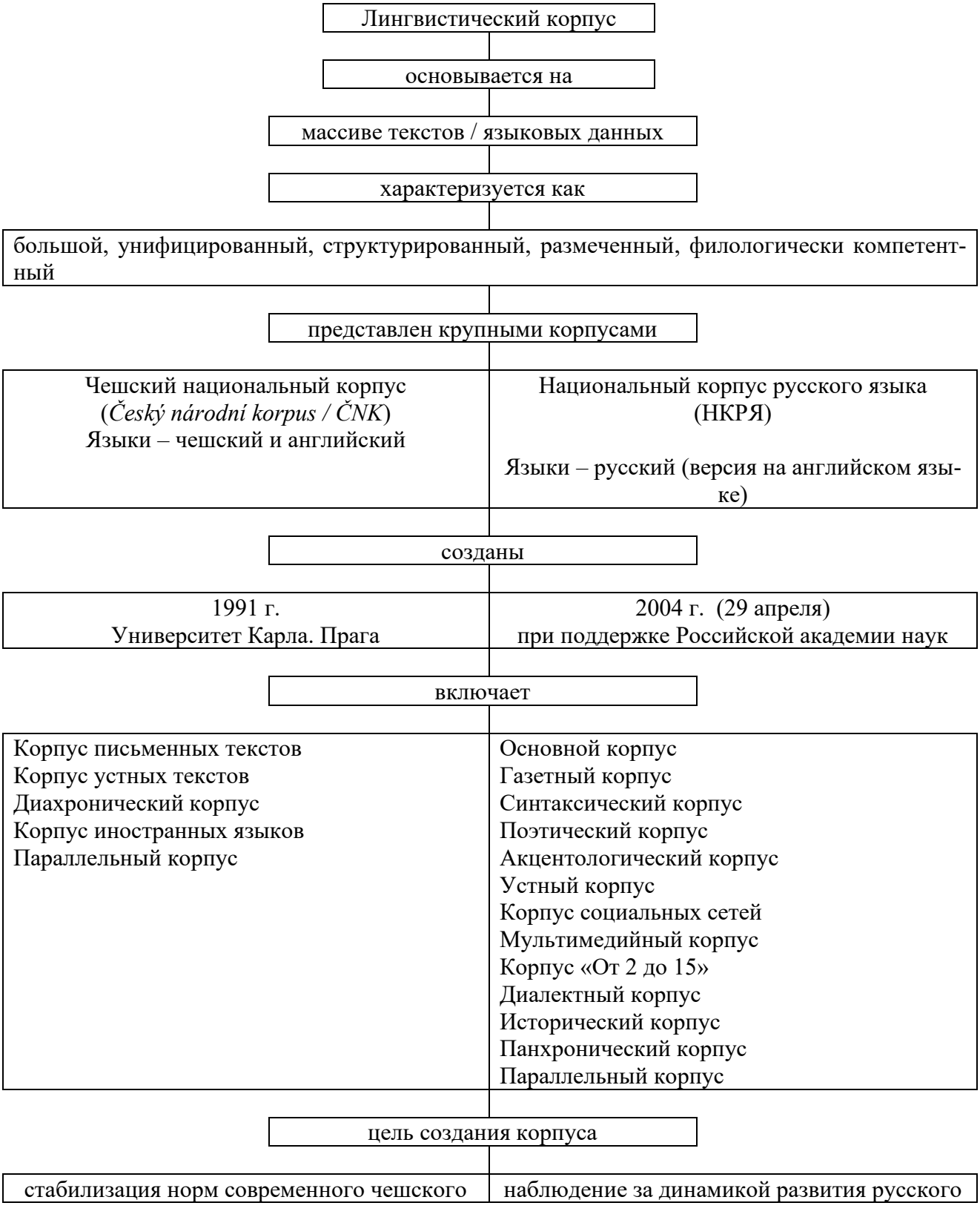
| **Критерии для сопоставления объектов** | **Национальный корпус русского языка** | **Фундаментальная электронная библиотека (ФЭБ)** |
|---|---|---|
| **Общее** | | |

| | | |
|---|---|---|
| 1. Единица хранения | собрание текстов | собрание текстов |
| 2. Формат хранения | электронный | электронный |
| 3. Степень доступности | открытый доступ пользователей к сетевой системе. | открытый доступ пользователей к сетевой системе. |
| **Различное** | | |
| 1. Цель создания | *собрание независимых корпусов текстов для решения определенных лингвистических задач.* | *создание многофункциональной программно-информационной среды для специалистов по русской филологии и фольклористике.* |
| 2. Назначение | *изучение текстов,*<br>*проведение лингвистических исследований по корпусу текстов.* | *чтение текстов,*<br>*проведение исследований с использованием полнотекстовой информационной системы,*<br>*распространение русской словесности и расширение международных культурных связей.* |
| 3. Обработка текстов | *обработка на техническом языке – разметка (собрание массива аннотированных /размеченных текстов).* | *обработка – собрание текстов в авторской редакции.* |
| 4. Язык объекта сопоставления | *русский язык и версия на английском языке* | *русский язык* |
| 4. Характерные признаки сопоставляемых объектов | *мультимедийность* | *энциклопедичность, иерархическая информационная структура* |
| 5. Типы собранных текстов | *художественная литература,*<br>*газетные тексты,*<br>*блогосфера,*<br>*устная речь,*<br>*диалектная речь,*<br>*просторечие,*<br>*устные тексты разных жанров,*<br>*научная речь,*<br>*аудио- и видеофайлы,*<br>*кино-, теле-, радио- и театральные постановки.* | *русская литература (XI–XX вв.) и фольклор,*<br>*история русской филологии и фольклористики,*<br>*архивные материалы, документы об авторе текстов,*<br>*мемуары,*<br>*филологические исследования.* |
| 6. Наличие справочной информации | *статистика (текстов, слов),*<br>*частотный словарь,*<br>*инструменты исследования (портреты корпуса, подкорпуса. слова).* | *научная литература,*<br>*словари и энциклопедии,*<br>*каталог ссылок на филологические и специализированные сайты.* |
| 7. Область применения | *специальные научные исследования,*<br>*преподавание русского языка, обучение русскому языку как иностранному.* | *гуманитарное образование всех уровней.* |
| 8. Пользователи | *ученые,*<br>*преподаватели,*<br>*учащиеся, студенты, аспи-* | *широкий круг пользователей* |

| | | |
|---|---|---|
| | *ранты.* | |

Б) Модель работы с денотатным графом
1. Представьте самые крупные корпуса в формате денотатного графа.

2. Обозначьте основные свойства Чешского национального корпуса (*Český národní korpus / ČNK*) и Национального корпуса русского языка (НКРЯ)
**ОТВЕТ:**



| | |
|---|---|
| языка,<br>нормализация политической ситуации (сотрудничество с международным научным сообществом) | языка |

**Круглый стол 1** (УК-4, ИД-УК-4.2, ИД-УК-4.3)

**УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.
**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.
**ИД-УК-4.3** Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств

Круглый стол на тему: «Корпусная лингвистика: исторический и лингводидактический аспекты»

Время выполнения 90 мин.
Проводится в подгруппах по 5-7 чел.
Форма работы – самостоятельная, групповая.
Проблема: новые технологии в лингвистике
Ожидаемый результат: Систематизация положений, обсуждаемых на семинаре

| № | Вопрос | Ответ | Компетенция |
|---|---|---|---|
| **УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.<br>**ИД-УК-4.1** Подготовка и редактирование различных академических текстов.<br>**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.<br>**ИД-УК-4.3** Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств. | | | |
| 1 | Создание и разметка корпусов текстов | **Лингвистическая разметка текстовых корпусов** бывает внешней (метаразметкой) и внутренней (собственно лингвистической), при этом синонимами выступают понятия *тэггинг* и *аннотирование.* Несмотря на то, что в иностранной литературе зачастую понятия *annotation*, *markup* и *tagging* используются взаимозаменяемо, именно аннотирование относится к внешней разметке. В литературе приводятся различные определения аннотирования. Аннотирование – это процесс составления аннотации, а сама аннотация – краткая характеристика документа, поясняющая его содержание, назначение, форму, другие особенности. Если семантика аннотации выражена явным образом, то такая аннотация – семантическая. Таким образом, процесс формирования списка ключевых слов является видом семантического аннотирования.<br>Разметка текста заключается в приписывании текстам и их компонентам дополнительной информации (метаданных). Метаданные можно поделить на 3 типа: экстралингвистические, относящиеся ко всему тексту; данные о структуре текста; лингвистические | УК-4,<br>ИД-УК-4.2,<br>ИД-УК-4.3 |

| | | | |
|---|---|---|---|
| | | метаданные, описывающие элементы текста. Метаописание текстов корпуса включает как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ). Эти данные обычно вводятся вручную. Структурная разметка документа (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. В Национальном корпусе русского языка в настоящее время используется пять типов разметки: метатекстовая, морфологическая (словоизменительная), синтаксическая, акцентная и семантическая. | |
| 2 | Словарь-конкорданс и его применение в рамках корпусной лингвистики | **Поиск в корпусе данных позволяет по выбранному** слову построить конкорданс. Конкорданс – это список всех употреблений определенного языкового выражения в контексте и со ссылкой на источник при необходимости. В данной коннотации этот термин часто применяется в корпусной лингвистике. В исходном своем толковании этот термин также используется для обозначения списка ключевых слов книги или работы, градированных в алфавитном порядке, с их контекстами.<br>Конкордансы часто используются в прикладной лингвистике: при переводе, в лексикографии, при обучении и изучении языка, при анализе текста.<br>Конкордансы применяются для – решения следующих лингвистических задач:<br>-создавать списки слов;<br>-сопоставлять разные использования одного слова;<br>-искать и исследовать фразы и идиомы;<br>-анализировать ключевые слова;<br>-искать перевод различных словосочетаний или слов;<br>-анализировать частоту употребления какого-либо слова и словосочетания.<br>Словарь-конкорданс является особым типом словаря, в котором каждое слово приводится с минимальным контекстом.<br>При изучение текста можно выделить поисковую, эвристическую, аналитическую функции конкорданса, а так же функции индексации и сравнения. *Поисковая функция* позволяет быстро находить нужный фрагмент текста, используя заданное слово или словосочетание. *Эвристическая функция* конкорданса отличает его от именных указателей наличием контекста. Зачастую, контексты позволяют лингвисту или простому читателю увидеть новую трактовку текста. *Аналитическая функция* позволяет проводить анализ различный языковых показателей, таких как лексемы, ключевые слова, частота их употребления в тексте и т.д. *Функция индексации* дает возможность при подготовке текста к публикации создавать индексы и списки слов. *Функция сравнения* применяется при сравнении в тексте всевозможных коннотаций и употреблений слова.<br>В отличие от словаря, который опирается на словарную статью, конкорданс приводит примеры – контексты словоупотреблений. Конкорданс – фун- | УК-4, ИД-УК-4.2, ИД-УК-4.3 |

| | | | |
|---|---|---|---|
| | | дамент для созданий разнообразных дифференциальных, в том числе и толковых, словарей. | |
| 3 | Метод автоматической кластеризации текстов и его применение. | **Основные процедуры обработки** естественного языка: токенизация, лемматизация, стемминг, парсинг.<br>*Токенизация* – разбиение потока символов в естественном языке на отдельные значимые единицы (токены, словоформы), является необходимым условием для дальнейшей обработки естественного языка. Если бы языки обладали совершенной пунктуацией, токенизация не представляла бы сложности – даже самая простая программа могла бы разделить текст на слова, руководствуясь пробелами и знаками препинания. Но в действительности языки подобной пунктуацией не обладают, что усложняет задачу токенизации. Например, в английском языке встречаются случаи, которые не могут быть однозначно токенизированы.<br>Другая специфическая задача морфологического анализа – это *лемматизация*, процесс образования первоначальной формы слова, исходя из других его словоформ. Во многих языках слово может встречаться в нескольких формах с различными флексиями. Базовая форма, зафиксированная в словаре, называется ***леммой*** слова. Лемматизация представляет собой процесс группировки различных флективных форм одного слова таким образом, чтобы при анализе они обрабатывались как одно слово. Лемматизация связана с идентификацией частей речи и включает в себя сокращение слов из корпуса до соответствующих им лексем. Именно лемматизация позволяет исследователю выделять и изучать все варианты отдельной лексемы без необходимости введения всех возможных вариантов.<br>Процесс, несколько отличный от лемматизации, называется *стеммингом*, он состоит в нахождении стема (основы) слова. Разница заключается в том, что стеммер обрабатывает отдельное слово без знания контекста, и, таким образом, не может дифференцировать слова, которые имеют разные значения в силу отнесенности к разным частям речи. Стеммеры обычно более просты для реализации и быстрее обрабатывают данные, а более низкая точность их работы может не иметь решающего значения для многих приложений. Лемма является базовой формой для токена, и это соответствие будет обнаружено как при стемминге, так и при лемматизации.<br>*Парсинг* – это процесс сопоставления линейной последовательности лексем (слов, токенов) языка с его формальной грамматикой. Результатом обычно является дерево зависимостей (синтаксическое дерево). Построение автоматических синтаксических анализаторов (парсеров) для больших корпусов является одной из самых важных областей компьютерной лингвистики. Большинство подходов объединяет качественные и количественные измерения. Наряду с разными статистическими подходами, которые тренируются на снабженных вручную пометами синтаксических деревьях, многие синтаксические анализаторы используют основанные на правилах | УК-4,<br>ИД-УК-4.2,<br>ИД-УК-4.3 |

| | | | |
|---|---|---|---|
| | | или основанные на ограничениях подходы, которые прямо моделируют специфические лингвистические теории. Разработка этих синтаксических анализаторов тесно переплетается с развитием этих теорий. Поскольку большинство предложений неоднозначны в любой теории, на основе правил (или перечня ограничений) должна быть разработана стратегия снятия неоднозначности. Многие стратегии снятия неоднозначности полагаются на количественные данные – частоту данной структуры в данном корпусе (тип), ограничения на выборку для данных лексических единиц, которые были получены или выделены из корпусных данных, и т.д. | |

**Круглый стол 2** (УК-4, ИД-УК-4.2, ИД-УК-4.3)

**УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.
**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.
**ИД-УК-4.3** Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств.

Круглый стол на тему: «Семантические исследования на материале корпусов текстов».

Время выполнения 90 мин.
Проводится в подгруппах по 5-7 чел.
Форма работы – самостоятельная, групповая.
Проблема: новые технологии в семантике
Ожидаемый результат: Систематизация положений, обсуждаемых на семинаре

| № | Вопрос | Ответ | Компетенция |
|---|---|---|---|
| **УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия. **ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации. **ИД-УК-4.3** Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств. | | | |
| 1 | Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы. | **Ценность корпуса** определяется глубиной разметки: каждому слову приписывается исчерпывающая морфологическая информация, а для каждого предложения строится полное синтаксическое дерево зависимостей. Разметка производится автоматически, однако работа над корпусом включает ручную правку всех предложений человеком. Работа по обогащению корпуса семантической информацией включает в себя четыре этапа: (1) разработку инвентаря семантических дескрипторов, (2) создание семантического словаря с приписанными лексемам семантическими дескрипторами и согласование этого словаря с комбинаторным словарем (КС) системы ЭТАП, (3) внедрение семантической информации в уже размеченный морфологически и синтаксически корпус текстов; (4) создание инструментария для | УК-4, ИД-УК-4.2, ИД-УК-4.3 |

| | | | |
|---|---|---|---|
| | | работы с семантической информацией.<br>Предлагаемый набор семантических дескрипторов (семантический метаязык) должен в конечном счете решать две задачи: 1. обеспечивать лингвистически содержательную классификацию всей лексики – и предметной, и предикатной; 2.,в соединении с морфологической и синтаксической разметкой текстов предоставлять исследователю существенную информацию о закономерностях поведения элементов различных лексико-семантических классов в текстах. В качестве дескрипторов везде, где это возможно, используются слова естественного языка. При разработке инвентаря семантических дескрипторов важно разделить лексемы на два основных типа – предметные (названия животных, птиц, рыб, овощей, фруктов, камней, гор, планет, светил и т. п.) и предикатные. Предметные и предикатные дескрипторы делятся на две подгруппы – родовые и видовые. Родовые дескрипторы обозначаются существительными (например, «животное», «совокупность», «состояние», «действие»), тогда как видовые – прилагательными («домашний», «природный», «речевой», «ментальный», «физический»). Предикатным словам, кроме родовых и видовых дескрипторов, приписываются семантические роли по каждой из валентностей. Например, глаголу вязать в значении «плести спицами или крючком» приписываются семантические роли «агенс» *(Маша вяжет),* «результат» *(вязать шарф*), «пациенс» *(вязать из шерсти)* и «инструмент» *(вязать крючком*). С учетом семантических ролей общий объем дескрипторов составляет 250–300 единиц. Предметной и предикатной лексике соответствуют две разные семантические классификации языковых единиц – таксономическая и фундаментальная. Предметные дескрипторы членят словарь не с научной, а с наивно-энциклопедической точки зрения.<br>Синтаксическая разметка является результатом *парсинга*, выполняемого на основе данных морфологического анализа. Этот вид разметки описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции (например, придаточное предложение, глагольное словосочетание и т.д.).<br>В отличие от морфологии, способы представления синтаксической структуры и синтаксических отношений не столь унифицированы. Наблюдается разнообразие синтаксических теорий и формализмов:<br>– грамматика зависимостей;<br>– грамматика непосредственно составляющих;<br>– грамматика структурных схем;<br>– традиционные синтаксические учения о членах предложения;<br>– функциональная грамматика;<br>– семантический синтаксис и др.<br>Синтаксический анализ для русского языка чаще всего представлен структурами зависимостей. | |
| 2 | Анализ семантических помет в Национальном корпусе русского языка. | **Национальный корпус русского языка снабжен разными видами разметки**, в том числе – морфологической и семантической. Семантическая разметка включает пометы таксономического класса, мереологию, оценку и др. типы информации. В разных видах | УК-4, ИД-УК-4.2, ИД-УК-4.3 |

| | | | |
|---|---|---|---|
| | | разметки есть разные виды неоднозначности. Что касается семантической разметки, то связанная с ней проблема неоднозначности состоит в следующем. В словаре у каждого значения многозначного слова есть своя собственная семантическая помета. Однако когда программа автоматически расставляет пометы в тексте, то она каждому вхождению слова приписывает все пометы, которые есть в словаре, потому что программа не знает, в каком значении выступает слово в данном тексте. Тогда нужно различить эти значения. Одним из эффективных путей решения этой проблемы являются семантические фильтры, т.е. семантические правила, которые позволяют оставлять при каждом вхождении слова только одну помету. Таким образом, многозначность снимается с точностью до семантического класса (т.е. с точностью до семантической пометы). Для снятия многозначности с помощью фильтров используется принцип контекстной однозначности. В словаре у слова может быть несколько значений, а в тексте (кроме специальных случаев языковой игры) – одно значение. Поскольку слово обычно не выступает в тексте изолированно, а включено в определенный контекст (в другой терминологии – в конструкцию), то этот контекст (конструкция) и является, в самом грубом и общем виде, фильтром. Фильтр работает следующим образом: формулируются семантические, морфологические и синтаксические условия, в которых реализуется некоторое значение слова; эти условия записываются в виде контекста; осуществляется поиск слова в заданном контексте; результатом этого поиска будет корпус примеров, где слово выступает в заданном значении, соответствующем определенной семантической помете. Остальные пометы стираются. Далее процедура повторяется для остальных семантически размеченных значений слова. | |
| 3 | Статистические исследования на материале корпусов текстов. | **Статистические методы** играют важную роль в анализе текстов в рамках корпусных исследований. Они позволяют исследователям получить количественные данные о различных языковых явлениях и провести объективный анализ текстов.<br>*Частотный анализ.* Один из основных статистических методов, используемых в анализе текстов, – это частотный анализ. Он позволяет определить частотность и распределение определенных слов, выражений или грамматических конструкций в тексте или в корпусе текстов.<br>Частотный анализ может быть полезен для выявления наиболее употребляемых слов или выражений в определенном жанре текстов, для сравнения частотности слов в разных языках или для изучения изменений в употреблении слов в течение времени.<br>*Коллокационный анализ.* Коллокационный анализ – это метод, который позволяет исследователям выявить наиболее типичные сочетания слов в тексте или в корпусе текстов. Он основан на статистическом анализе совместной встречаемости слов и позволяет выявить семантические и грамматические связи между словами. Коллокационный анализ может быть полезен для изучения лексических особенностей языка, выявления идиоматических выражений или для определения наиболее типичных словосоче- | УК-4, ИД-УК-4.2, ИД-УК-4.3 |

| | | | |
|---|---|---|---|
| | | таний в определенном жанре текстов.<br>*Статистический анализ стилей и жанров.* Статистические методы также могут быть использованы для анализа стилей и жанров текстов. Исследователи могут проводить статистический анализ различных лингвистических признаков, таких как длина предложений, использование определенных частей речи или грамматических конструкций, чтобы выявить характерные особенности стиля или жанра. Статистический анализ стилей и жанров может быть полезен для автоматической классификации текстов по жанрам или для выявления стилистических особенностей в текстах разных авторов.<br>*Машинное обучение и анализ текстов.* Статистические методы также широко применяются в машинном обучении и анализе текстов. Используя статистические модели и алгоритмы, исследователи могут разрабатывать системы автоматического распознавания и классификации текстов, а также системы автоматического извлечения информации из текстов. Машинное обучение и анализ текстов на основе статистических методов могут быть полезными для обработки больших объемов текстов, автоматического анализа семантики текстов или для разработки систем машинного перевода. | |

**Устный опрос** (УК-4, ИД-УК-4.1, ИД-УК-4.2, ИД-УК-4.3)

**УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.
**ИД-УК-4.1** Подготовка и редактирование различных академических текстов.
**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.
**ИД-УК-4.3** Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств.

**Устный опрос по вопросам:**

Форма работы – самостоятельная, индивидуальная

| № | Вопрос | Ответ | Компетенция |
|---|---|---|---|
| **УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.<br>**ИД-УК-4.1** Подготовка и редактирование различных академических текстов. | | | |
| 1 | Что может являться единицей корпуса? | Корпус текстов - это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то полные фрагменты макроструктуры текстов данной проблемной области. Корпус текстов характеризуется четырьмя основными параметрами: 1. он должен быть достаточно большого объема; 2. корпус должен быть структурированным или размеченным; 3. тексты, составляющие определенного кор- | УК -4<br>ИД-УК-4.1 |

<table>
<tr><td></td><td></td><td>пуса, должны быть в электронном варианте; 4. в понятие «Электронный корпус» входит, как правило, специальное программное обеспечение для работы с этим корпусом.</td><td></td></tr>
<tr><td colspan="4">УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.<br>ИД-УК-4.2 Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.</td></tr>
<tr><td>2</td><td>Определите виды корпуса по указанному признаку</td><td><table>
<tr><td>Признак</td><td>Виды корпуса (ответ)</td></tr>
<tr><td>Форма хранения</td><td>звуковые, письменные, смешанные</td></tr>
<tr><td>Язык текстов</td><td>русский, английский</td></tr>
<tr><td>Параллельность</td><td>одноязычные, двуязычные, многоязычные</td></tr>
<tr><td>Стиль</td><td>литературные, диалектные, разговорные, публицистические, терминологические, смешанные</td></tr>
<tr><td>Способ доступа</td><td>свободно доступные, коммерческие, закрытые</td></tr>
<tr><td>Разметка</td><td>размеченные, неразмеченные</td></tr>
<tr><td>Характер разметки</td><td>морфологические, синтаксические, семантические, просодические</td></tr>
</table></td><td>ИД-УК-4.1<br>ИД-УК-4.2</td></tr>
<tr><td colspan="4">УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.<br>ИД-УК-4.1 Подготовка и редактирование различных академических текстов.<br>ИД-УК-4.2 Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.<br>ИД-УК-4.3 Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств.</td></tr>
<tr><td>3</td><td>Поясните, что означают следующие понятия: «исследовательский корпус»</td><td>Исследовательский корпус – это корпус, предназначенный для изучения раз¬личных аспектов функционирования языковой системы. Они строятся не post factum — после проведения какого-либо исследования, а до его проведения. Этот тип корпусов данных ориентирован на широкий класс лингвистических задач.</td><td>УК -4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3</td></tr>
<tr><td>4</td><td>Поясните, что означают следующие понятия: «статический корпус»</td><td>Статический корпус<br>Первоначально корпусы текстов создавались как статические образования, отражающие определенное временное состояние языковой системы. Типичными представителями этого вида корпусов – авторские</td><td>УК -4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3</td></tr>
</table>

| | | | |
|---|---|---|---|
| | | корпусы, коллекции текстов писателей. | |
| 5 | Поясните, что означают следующие понятия: «параллельный корпус» | **Параллельный корпус**<br>1) корпусы, представляющие множество текстов-оригиналов, написанных на какомлибо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков;<br>2) корпусы, объединяющие тексты из одной и той же тематической области, независимо написанные на двух или нескольких языках. | УК -4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 6 | Как можно использовать рассмотренные корпусы в лингвистическом исследовании / в практике перевода? | *Корпусы параллельных текстов* формируются для научных и практических целей (в частности, для преподавания иностранных языков). По своей структуре это подмножество текстов на языке-источнике и одно или несколько подмножеств текстов, которые являются переводами текстов языка-источника на языки-цели. Например, английский текст «Alice in Wonderland» и его переводы на немецкий, французский и русский языки могут формировать такой корпус или быть частью большего корпуса параллельных текстов. Корпусные методы исследования перевода привнесены в область аудиовизуального перевода, применяются возможности корпусных менеджеров в дескриптивном переводоведении, в преподавании языка в иностранной аудитории. Учебные тексты позволяют классифицировать типы ошибок и учитывать их в процессе преподавания. Сведения такого рода учитываются в англоязычных учебных словарях.<br>*Исследовательский корпус* может быть использован как инструмент научного исследования, в грамматических и лексикологических исследованиях, в лингвистической, судебно-лингвистической экспертизе, исследование частотных характеристик лексем (лемм), исследование коллокаций (сочетаний лексем).<br>*Статический корпус* необходим для исследования нормы / узуса, в социолингвистических исследованиях, для изучения устной речи. Наличие электронных текстов, принадлежащих одному автору, дает возможность расширить круг задач, традиционно решаемых стилистикой и авторской стилеметрией. | УК -4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |

**Тестирование** (УК-4, ИД-УК-4.1, ИД-УК-4.2)

**УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.
**ИД-УК-4.1** Подготовка и редактирование различных академических текстов.
**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.

**Тест**
Время выполнения 15 мин.
Количество вопросов 10.
Форма работы – самостоятельная, индивидуальная.
Способ проведения теста: бланковый

***Инструкция для тестируемых:***
Внимательно читать задания к тестовым вопросам и четко отвечать на них в зависимости от типа задания. Тест выполняется самостоятельно в течение 15 минут.
Номинальная шкала предполагает, что за правильный ответ к каждому заданию выставляется один балл, за не правильный – ноль. В соответствии с номинальной шкалой, оценивается всё задание в целом, а не какая-либо из его частей.
Процентное соотношение баллов и оценок по пятибалльной системе:
«2» - равно или менее 54%
«3» - 55% - 69%
«4» - 70% - 84%
«5» - 85% - 100%.

***Инструкция для проверяющих:***
Правильные ответы тестовых заданий выделены. Номинальная шкала предполагает, что за правильный ответ к каждому заданию выставляется один балл, за не правильный — ноль. В соответствии с номинальной шкалой, оценивается всё задание в целом, а не какая-либо из его частей.
Процентное соотношение баллов и оценок по пятибалльной системе:
«2» - равно или менее 54%
«3» - 55% - 69%
«4» - 70% - 84%
«5» - 85% - 100%.

**Тест с ответами**

| № | Вопрос | Ответ | Компетенция |
|---|---|---|---|
| **УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.<br>**ИД-УК-4.1** Подготовка и редактирование различных академических текстов.<br>**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации. | | | |
| 1 | **1. Корпусная лингвистика занимается**<br>а) объяснением фактов языка;<br>**б) объяснением фактов речи;**<br>в) изучением системности и структурности языка; | | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |

| | | |
|---|---|---|
| | г) типологией языковых ситуаций. | |
| 2 | **2. Машинный перевод является предметом описания**<br>а) структурной лингвистики;<br>**б) компьютерной лингвистики**;<br>в) корпусной лингвистики;<br>г) социолингвистики. | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |
| 3 | **Какой из представленных корпусов реально не существует?**<br>а) Мангеймский корпус немецкого языка;<br>б) Национальный корпус русского языка;<br>в) Упсальский корпус русского языка;<br>**г) Национальный корпус французского языка;**<br>д) Корпус современного американского английского языка. | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |
| 4 | **Лингвистический корпус – это**<br>а) собрание текстов, сборники, которые носят довольно случайный характер;<br>б) вид издания, которое концентрирует в однотипно оформленных единицах (томах) все или основные произведения одного автора в качестве его научного, литературно-художественного или публицистического наследия;<br>**в) тексты, собранные и размеченные по определённому стандарту и обеспеченные специализированной поисковой системой;**<br>г) собрание письменных сообщений, объективированных в виде письменных документов, имеющих прагматическую установку и соответственно литературно обработанных;<br>**д) информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме, представляет данный язык на определенном этапе (или этапах) его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов.** | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |
| 5 | **Выберите критерии, по которым могут выделяться типы корпусов текстов:**<br>**а) по языку представления текстов**<br>б) по статистическому критерию;<br>**в) по жанровой принадлежности;**<br>**г) по форме хранения;**<br>д) по удобству чтения;<br>е) по возможностям читательского восприятия. | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |
| 6 | **Цель создания корпуса –**<br>а) для отражения художественного содержания текстов;<br>б) для работы в области машинного перевода;<br>**в) для научных исследований и обучения языку;**<br>г) для создания образовательной среды, объединяющей литературные произведения в единую систему. | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |
| 7 | **Найдите среди представленных параметров типы разметок корпуса:**<br>а) метаязыковой;<br>**б) семантический;**<br>в) гносеологический;<br>**г) морфологический.** | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |
| 8 | **Какие из представленных единиц являются подкорпусами Национального корпуса русского языка:**<br>**а) параллельный корпус;**<br>**б) обучающий корпус русского языка;**<br>в) стратификационный корпус;<br>**г) газетный корпус;**<br>**д) спряжение глаголов;**<br>е) динамический корпус;<br>ж) речевой корпус. | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2 |
| 9 | **Разбиение на орфографические слова в корпусе – это**<br>а) конкорданс; | УК-4<br>ИД-УК-4.1 |

<table>
<tr><td></td><td>б) токенизация;<br>в) лемматизация;<br>г) кластеризация.</td><td>ИД-УК-4.2</td></tr>
<tr><td>10</td><td>Найдите ошибку в сопоставлении традиционной и корпусной лингвистики
<table>
<tr><th></th><th>Корпусная лингвистика</th><th>Традиционная лингвистика</th></tr>
<tr><td>а)</td><td>Текст рассматривается как некоторая физическая сущность.</td><td>Текст рассматривается как некоторая абстракция.</td></tr>
<tr><td>б)</td><td>Составление грамматики конкретных языков.</td><td>Изучает языковые универсалии</td></tr>
<tr><td>в)</td><td>Основное внимание уделяется форме.</td><td>Основное внимание не только форме, но и содержанию.</td></tr>
<tr><td>г)</td><td>Предпочитаются искусственные примеры, из изолированных от текста словоупотреблений.</td><td>Проводится работа с лингвистическими данными (словоупотреблениями) в том виде, в каком они встречались в контексте.</td></tr>
</table></td><td>УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2</td></tr>
</table>

**4.2.** Оценочные материалы **промежуточного контроля** успеваемости по учебной дисциплине, в том числе самостоятельной работы обучающегося, типовые задания

**Зачет с оценкой** (УК-4, ИД-УК-4.1, ИД-УК-4.2, ИД-УК-4.3)

**УК-4** Способен применять современные коммуникативные технологии, в том числе на иностранном (ых) языке (ах), для академического и профессионального взаимодействия.
**ИД-УК-4.1** Подготовка и редактирование различных академических текстов.
**ИД-УК-4.2** Готовность к участию в профессиональных дискуссиях и грамотное использование деловой, устной и письменной коммуникации.
**ИД-УК-4.3** Навыки межличностного делового общения, в том числе на иностранных языках с применением профессиональных языковых форм и средств.

**Устный опрос по вопросам:**
Время на подготовку 20 мин

Перечень теоретических вопросов к зачету:

| Вопрос | Ответ | Компетенции |
|---|---|---|
| 1. Компьютерная лингвистика как направление научной деятельности, ее цели и задачи. | *Компьютерная лингвистика – направление в прикладной лингвистике, ориентированное на использование программ, компьютерных технологий организации и обработки данных для моделирования функционирования языка. создание и использование электронных корпусов текстов; создание электронных словарей, тезаурусов; автоматический перевод текстов; автореферирование; автоматическое распознавание речи и автоматический синтез речи.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 2. Корпусная лингвистика как новое | *Корпусная лингвистика – это раздел прикладной лингвистики, занимающийся* | УК-4<br>ИД-УК-4.1 |

| | | |
|---|---|---|
| направление в языкознании. | *разработкой общих принципов построения и использованием лингвистических корпусов. Корпус – это информационно-справочная система (специализированная поисковая система), собрание снабженных стандартной разметкой текстов на определенном языке в электронной форме.* | ИД-УК-4.2<br>ИД-УК-4.3 |
| 3. Критерии создания корпусов | *Стратегия построения корпуса.* ***Репрезентативность*** *(представительность) – способность корпуса текстов отражать все свойства проблемной области. Репрезентативность корпуса указывает на то, что единицы проблемной области отражаются пропорционально в корпусе данных.* ***Полнота*** *требует учета релевантных явлений, даже если это не соответствует идее пропорционального сужения.* ***Экономичность*** *корпуса текстов – экономия усилий исследователя при изучении проблемной области. Чем более «экономичен» корпус, тем выше порог отображения.* ***Структуризация материала*** *– опись данных корпуса, в которой единицы хранения характеризуются по тем параметрам, которые могут оказаться важными для пользователя.* ***Компьютерная поддержка*** *– комплекс программ по обработке данных, обеспечивающих функции составления конкордансов, статистической инвентаризации, автоматической словарной обработки, лемматизации.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 4. Корпусная лингвистика и проблемы перевода | *Создание автоматизированных лексикографических систем в помощь переводчику. Создание, ведение и издание переводных словарей. Задачи – выбор принципов лексикографирования и создания специального «словаря для переводчика» (определение макроструктуры словаря); выбор формата представления информации в словаре (определение микроструктуры словаря); выбор принципов и методов формирования и использования корпуса текстов. Тексты оригинала и перевода являются параллельными или битекстами.* ***Параллельный корпус текстов*** *– совокупность документов, переведенных на два или более языков, выровненных по предложениям и размеченных, написанных на одну тему и на одном языке авторами с разными родными языками. Решение проблем перевода с использованием корпуса: интерпретация оригинала и текстопостроения перевода; определение лексической сочетаемости; сохранение стилистического рисунка исходного текста, правильность выбранных грамматических и синтаксических конструкций; сокращение затрачиваемого времени на перевод.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 5. Британские корпуса и словари. | *Британский национальный корпус (British National Corpus). Время создания – 1990-е гг. Объём этого знаменитого корпуса составляет 100 миллионов словоупотреблений. BNC – эталон лингвистического корпуса. Содержит образцы письменного и разговорного британского английского языка из широко-* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |

| | | |
|---|---|---|
| | *го круга источников. Корпус охватывает британский английский конца XX века, представленный широким разнообразием жанров, и задуман как образец типичного разговорного и письменного британского английского языка того времени. BNC является одноязычным корпусом, так как он содержит образцы только британского английского языка, хотя иногда в текстах встречаются слова и фразы из других языков. Это синхронический корпус, так как в нём содержатся примеры использования языка только одного временного периода – конец XX века. По этой причине BNC не может служить источником данных об истории развитии британского варианта английского языка. BNC сбалансированный корпус, включает данные из различных источников. Корпус BNC содержит частеречную разметку. Включает письменный и разговорный подкорпусы. Недостаток корпуса – слишком общая классификация текстов.* | |
| 6. Корпуса американского варианта английского языка. | *Корпус современного американского английского (Corpus of Contemporary American English, COCA) – электронный корпус текстов, созданный профессором корпусной лингвистики Марком Дэвисом из Университета Бригама Янга в 2000–2003 годах на основе текстов журнала Time, написанных с 1923 года. Это наибольший (450 млн слов) корпус текстов американского варианта английского языка. Свободно доступный корпус, включающий большое разнообразие текстов различных жанров. Он составлен из более чем 160 тыс. текстов, включая по 20 млн слов за каждый год с 1990 по 2011. Это наиболее широко используемый структурированный корпус текстов, ежемесячно его используют примерно 10 000 человек. Архитектура корпуса – принцип других проектов Марка Дэвиса. Существует центральная база n-граммов, которая содержит информацию о каждом из ста миллионов слов корпуса. Они связаны с таблицами, позволяющими анализировать регистр, а также с отдельными таблицами для синонимов, лемм и форм, появляющихся у слова с течением времени. Характерная особенность корпуса – быстрый поиск (занимает менее секунды даже для самых сложных запросов, содержащих словоформу, часть речи, частоту и регистр). Корпус позволяет исследовать: изменение частоты и контекста использования слов и фраз, связанных с переменами в культурной и социальной жизни общества; языковые перемены в морфологии и грамматических конструкциях, колебания частоты использования тех или иных групп слов со временем, семантические изменения слов на протяжении XX века.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 7. НКРЯ как инструмент семантико-грамматического исследования лек- | *Национальный корпус русского языка (НКРЯ, Корпус) – собрание независимых корпусов,* | УК-4<br>ИД-УК-4.1 |

| | | |
|---|---|---|
| сики. | *основной инструмент поиска при лингвистических исследованиях. Корпусы большие по объёму и представительные – ценный материал для количественных и качественных исследований. Структура корпуса – основной (тексты 18-21 вв.), газетный, поэтический, акцентологический, устный корпус, корпус социальных сетей, мультимедийный, синтаксическом, обучающий, исторический, панхронический, диалектный, параллельный корпуса, корпус «От 2 до 15», корпус региональной прессы. Объём основного корпуса на декабрь 2023 года – 375 млн словоупотреблений, а общий объём корпусов – больше 2 млрд словоупотреблений. Тексты снабжены метаразметкой (по дате создания, автору, жанру и тому подобному); словоформы в текстах снабжены автоматической морфологической и семантической разметкой; параллельные тексты выровнены; тексты поэтического корпуса снабжены также особой метрической разметкой.* | ИД-УК-4.2<br>ИД-УК-4.3 |
| 8. Статистическое описание лексики в писательской лексикографии. | *Приход больших данных в лингвистику, развитие нейронных сетей позволяют по-новому исследовать материал по авторской лексикографии. Компьютеризация лексикографической деятельности заключается прежде всего в создании специализированных машинных банков данных и в разработке методов формирования этих банков, представления информации в банках и её использовании. На этой основе формируется целое новое направление лингвистики и лексикографии – корпусная лингвистика и лексикография. Ряд специализированных проектов ориентирован на исследование сочетаемости ключевых слов на материале русского языка с привлечением статистического аппарата. Таким ресурсом является система CoCoCo (“Collocations, Colligations, Constructions”), которая разрабатывается под руководством М. Копотева в Хельсинкском университете, представляет информацию о многословных выражениях на основе Национального корпуса русского языка. Многословные выражения понимаются предельно широко: к ним относятся идиомы, составные единицы, коллокации и коллигации. В системе предусмотрен поиск по леммам или токенам (словоформам), который можно ограничить морфологическими параметрами. При выборе части речи есть возможность задать значения грамматических категорий. Результаты поиска демонстрируют не все возможные словосочетания, а только наиболее значимые. Подобная информация основывается на применении статистических метрик. Для решения этой задачи используется расстояние Кульбака–Лейблера (мера KLD), которое позволяет оценить, какие грамматические признаки оказываются наиболее употребительными для данной лексической единицы.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |

| | | |
|---|---|---|
| 9. Автоматический семантический анализ текста на русском языке. | *Семантический анализ текста – наиболее сложная проблема в области искусственного интеллекта и компьютерной лингвистика. Главная проблема заключается в том, как «научить» компьютер однозначно верно трактовать образы, которые пытался передать автор текста. Цель семантического (смыслового) анализа – оценивание смысла передаваемой информации, соотношения ее с информацией, которая хранилась до появления обрабатываемой информации. Семантические связи между словами или другими единицами языка отражаются в семантических словарях. Задачами семантического анализа являются: 1) построение семантической интерпретации слов и конструкций; 2) установление семантических отношений между различными элементами текста. При семантическом анализе предложений используют падежные грамматики и семантические валентности, а семантика предложения задается через связи главного слова (глагола) с его семантическими актантами. Основой семантического анализа – утверждение, что конкретное значение слова не является элементарной семантической единицей. Оно делится на более мелкие единицы – единицы словаря семантического языка, являющиеся своеобразными атомами, комбинации которых складываются в «молекулы» – значения слов естественного языка. Именно семантический анализ дает возможность решить проблемы многозначности (омонимии), которая часто возникает при автоматическом анализе на разных языковых уровнях. В заключении стоит отметить, что ценность автоматического анализа текста на данный момент особенно высока, поскольку человек уже не в состоянии самостоятельно обработать современные объемы информации. Автоматический анализ текста находит применение в филологии (определение авторства произведений, авторского стиля), в экспертных системах.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 10.Исходные понятия корпусной лингвистики: проблемная область, корпус данных, корпус текстов. | *Проблемная область – область реализаций языковой системы, содержащая феномены, подлежащие лингвистическому описанию, она имеет два измерения – языковое и речевое. В корпусной лингвистике языковой аспект фактически игнорируется, поскольку изначально фиксируется область привлекаемых данных – реализаций языковой системы. Проблемная область для разработчика – корпус как множество данных, обработка которых затруднена из-за того, что языковых реализаций слишком много. Корпус данных – сформированная по определенным правилам выборка данных из проблемной области, результат отображения из проблемной области. В отличие от проблемной области, корпус данных имеет только одно измерение – речевое. Лингвисту приходится по отдель-* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |

| | | |
|---|---|---|
| | *ным результатам деятельности языка делать выводы о функционировании языка как целого, как системы. Корпус текстов – это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, полные фрагменты макроструктуры текстов данной проблемной области. Типы корпусов данных – исследовательские корпусы (для изучения различных аспектов функционирования языковой системы), иллюстративные корпусы (для подтверждения и обоснования уже полученные результаты), динамические (функционирования языковых феноменов на временной шкале) и статические корпусы текстов (временное состояние языковой системы – авторские корпусы).* | |
| 11.Параллельный многоязычный корпус текстов, его структура и сфера применения. | *Параллельный корпус (Parallel Corpora) – электронный аналог параллельных переводных текстов, состоящий из множества блоков «текст-оригинал и один/несколько его переводов». Электронные тексты в корпусе могут представлять собой целое оригинальное словесное произведение или какую-либо его часть.*<br>*В современной корпусной лингвистике выделяется два вида параллельных корпусов: 1) многоязычный, или Comparable (Multilingual) Corpora, 2) переводной, или Translation Corpora.*<br>*Структурная организация корпуса: в виде традиционного текста со ссылкой на перевод/ы, в табличной «зеркальной» форме, что более удобно для восприятия и сравнения, в виде базы данных. Параллельные корпуса – прикладной потенциал в методике обучения иностранным языкам и переводу, а также в компьютерной лингвистике. Переводчику необходимы ресурсы-эталоны перевода и оценка перевода в «стандартных» условиях. Электронные параллельные корпусы и лингвистические компьютерные технологии позволяют значительно сократить временные затраты и предоставляют образцы профессионального перевода при изучении приемов и способов перевода.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 12.Аннотированные корпусы текстов, автоматизация их создания и коррекции. | *Большие текстовые корпуса давно и плодотворно используются в компьютерной лингвистике. Уровни аннотации текста: 1) лематизированные тексты, в которых для каждого слова указывается его основная форма и часть речи; 2) тексты с морфологической информацией, в которых для каждого слова указываются его основная форма, часть речи и полный набор морфологических характеристик; 3) тексты с синтаксической информацией, в которых для каждого слова указываются его основная форма, часть речи и морфологические характеристики, и для каждого предложения указывается его синтаксическая структура. Построение аннотированного корпуса осуществляется в полу-* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |

| | | |
|---|---|---|
| | *автоматическом режиме: аннотация вначале порождается системой автоматического морфологического и синтаксического анализа, а затем корректируется специалистом-лингвистом. Степень участия лингвиста в процессе аннотации определяется самим лингвистом в зависимости от сложности структуры текста. Типы лингвистической информации на каждом уровне: морфологическая информация – морфологические характеристики слова (часть речи, одушевленность, род, падеж, число, падеж, степень сравнения, краткость, репрезентация, вид, время, лицо, залог); синтаксическая информация – число узлов в синтаксической структуре равно числу слов в предложении. Инвентарь синтаксических отношений – деление на 6 больших групп: 1) актантные; 2) атрибутивные; 3) количественные; 4) обстоятельственные; 5) сочинительные; 6) служебные. Поиск в корпусе другой информации: распределение по годам, частота употребления слова по годам; статистика (метаатрибуты: автор, пол автора, сфера функционирования, тип текста, тематика текста, жанр).* | |
| 13.Опыт разработки корпусов текстов в России и за рубежом. | *руководством У. Фрэнсиса существует в компьютерном варианте и на микрофишах. Объем корпуса около 1 млн словоупотреблений. Корпус состоит из 500 текстов, каждый из которых включает 2 000 словоупотреблений). Ланкастерско-Осло-Бергенский корпус (LOB), британский вариант английского языка. Аннотированная версия корпуса LOB появилась в 1985 г. Лондонско-Лундский корпус (фиксация особенности грамматической системы английского языка в речи взрослого образованного носителя; 1960-е г. под руководством Рэндола Квирка в Лондонском университетском колледже; объем корпуса – 1 млн словоупотреблений, письменные текст и тексты устной речи). Бирмингемский корпус (мониторный / динамический корпус, автор проекта – Дж. Синклер, 7,3 млн словоупотреблений, 6 млн письменные тексты, 1,3 млн – устные тексты). Разработка Германии проекта LIMAS-корпуса (система немецко-английского машинного перевода). Корпусы текстов немецкой разговорной речи. «Корпус базового немецкого» (1961 г. в Стэнфорде). Корпусная лингвистика во Франции («Сокровищница французского языка»).* ***Корпусы текстов по русскому языку*** *(начало 1970-х годов). «Уппсальский машинный фонд русского языка», создававшийся с 1987 г. в Уппсальском университете. Собственно российский опыт составления корпусов: Корпус по дискурсивным словам русского языка, Корпус словаря языка Достоевского, Корпус текстов словаря языка Достоевского, Динамический корпус текстов по современной публицистике (90-е гг.), Динами-* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |

| | | |
|---|---|---|
| | *ческий корпус текстов как новая технология прикладной лингвистики.* | |
| 14.Практическое использование аннотированных корпусов текстов в системах автоматической обработки текстов. | *Аннотирование текстов – это процесс добавления метаданных и лингвистической информации к текстам в корпусе. Для эффективного поиска и анализа текстов в корпусе необходимо создать поисковые индексы. Аннотирования корпусов требует тщательной работы и экспертизы в области лингвистики, разработки интерфейса для взаимодействия пользователей с корпусом.* ***Примеры практического применения корпусов в лингвистике.*** *Компьютерная обработка текста – преобразование текста на искусственном или естественном языке с помощью компьютера. Системное программирование – создание программного обеспечения функционирования компьютера и работы пользователей. Издательское дело – одно из направлений автоматизации редакционно-издательских процессов. Автоматизированное редактирование – внесение в текст, находящийся в памяти компьютера, исправлений и дополнений. Эти тенденция прогнозируется и прослеживается на примере развития АОТ-систем (АОТ – автоматическая обработка текста), представляющих коммерческий интерес. Решение следующих прикладных задач – машинного перевода, генерации текста, локализации и интернационализации, работа на ограниченном языке, создание текстовых документов (ввод, редактирование, исправление ошибок), информационный поиск и связанные с ним задачи.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |
| 15.Корпус и междисциплинарные исследования | *Междисциплинарное исследование – исследование одного объекта методами различных дисциплин: методология нескольких дисциплин – базовая, а другие – новые, инновационными по отношению к исследуемому объекту. Практическая целесообразность междисциплинарного исследования в корпусной лингвистике: корпусы текстов – продукт синергии а) лингвистических дисциплин, б) разработок в области машинной обработки языка, основанных на статистических и кибернетических методах, в) собственно информационных технологий. Исследования в области корпусной лингвистики – это способы и результаты усовершенствования подборки и управления корпусным материалом и усовершенствование имеющихся и разработка новых способов синтаксического, семантического и прочих видов анализа. Важным компонентом исследования является прагматика корпуса текстов – взаимодействие с пользователем. Два направления междисциплинарных исследований на основе корпусной лингвистики – фундаментальный (понимание), тяготеющие к психологии, и прикладной (внешнее моделирование полезных свойств), близкие к когнитивной науке.* | УК-4<br>ИД-УК-4.1<br>ИД-УК-4.2<br>ИД-УК-4.3 |

**ЛИСТ УЧЕТА ОБНОВЛЕНИЙ ОЦЕНОЧНЫХ СРЕДСТВ УЧЕБНОЙ ДИСЦИПЛИНЫ**

В оценочные средства учебной дисциплины внесены *изменения/обновления*, утверждены на заседании кафедры:

| № пп | год обновления оценочных средств | номер протокола и дата заседания кафедры |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |