

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Нижегородский государственный лингвистический университет имени
Н.А. Добролюбова»



УТВЕРЖДАЮ

Проректор по учебной
деятельности

Р.И. Иванова

29 августа 2023 г.

Компьютерная лингвистика. Введение. Языковые корпусы. Автоматизированный перевод (онлайн-курс)
(наименование дисциплины)

дополнительная профессиональная программа повышения квалификации

Форма обучения **заочная**

Общая трудоемкость **2 ЗЕТ**

Часов по учебному плану **72**
в том числе:
аудиторные занятия **18**
самостоятельная работа **53,95**
часов на контроль **0,05**

Виды контроля:

зачет

Распределение часов дисциплины по семестрам

Вид занятий	
Лекции	18
Практические (в том числе интеракт.)	
Семинарские (в том числе интеракт.)	
Итого ауд.	18
Часы на контроль	0,05
Контактная работа	18
Самостоятельная работа	53,95
Итого	72

Программу составил(и):
уч. степень, уч. звание, Ф.И.О. канд. филол. наук, доц. М.Б. Чиков

Рабочая программа
«Компьютерная лингвистика. Введение. Языковые корпуса. Автоматизированный перевод» (онлайн-курс)

разработана в соответствии с ФГОС:

Федеральный государственный образовательный стандарт высшего образования по направлению подготовки 45.03.02 Лингвистика, утвержденный приказом МИНОБРНАУКИ РОССИИ от 12 августа 2020 г., № 969.

Рабочая программа одобрена на заседании кафедры
романо-германских языков, перевода, зарубежной литературы и межкультурной коммуникации

Протокол от 29 августа 2023__ г. № 1

Зав. кафедрой канд. филол. наук Е.А. Сакулина
(*уч. степень, уч. звание, Ф.И.О.*)

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ КУРСА	
1.1	Цель освоения дисциплины: Сформировать сквозные цифровые компетенции у студентов гуманитарных специальностей, направлений и профилей подготовки
1.2	Задачи освоения дисциплины:
1.2.1	Дать представление о базовых инструментах цифровой обработки естественного языка
1.2.2	Дать представление о программных средствах обработки естественного языка
1.2.3	Дать представление о научно-исследовательских задачах, решаемых при помощи цифровых инструментов обработки естественного языка

2. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ КУРСА	
<i>ОПК-5. Способен работать с компьютером как средством получения, обработки и управления информацией для решения профессиональных задач.</i>	
Знать:	
Уровень Пороговый	На пороговом уровне: принципы междисциплинарного взаимодействия цифровых технологий и лингвистики, принципы функционирования компьютерных программ и сетевых ресурсов для обработки естественного языка, возможности применения цифровых технологий для решения учебных и научно-исследовательских лингвистических задач
Уровень Высокий	На высоком уровне: принципы междисциплинарного взаимодействия цифровых технологий и лингвистики, принципы функционирования компьютерных программ и сетевых ресурсов для обработки естественного языка, возможности применения цифровых технологий для решения учебных и научно-исследовательских лингвистических задач
Уровень Повышенный	На повышенном уровне: принципы междисциплинарного взаимодействия цифровых технологий и лингвистики, принципы функционирования компьютерных программ и сетевых ресурсов для обработки естественного языка, возможности применения цифровых технологий для решения учебных и научно-исследовательских лингвистических задач
Уметь:	
Уровень Пороговый	На пороговом уровне: ставить научно-исследовательские и учебные лингвистические задачи, решаемые с использованием цифровых технологий, и применять цифровые технологии для решения тех или иных поставленных задач
Уровень Высокий	На высоком уровне: ставить научно-исследовательские и учебные лингвистические задачи, решаемые с использованием цифровых технологий, и применять цифровые технологии для решения тех или иных поставленных задач
Уровень Повышенный	На повышенном уровне: ставить научно-исследовательские и учебные лингвистические задачи, решаемые с использованием цифровых технологий, и применять цифровые технологии для решения тех или иных поставленных задач
Владеть:	
Уровень Пороговый	На пороговом уровне: основными компьютерными программами и сетевыми ресурсами для обработки естественного языка, их функциональным инструментарием
Уровень Высокий	На высоком уровне: основными компьютерными программами и сетевыми ресурсами для обработки естественного языка, их функциональным инструментарием
Уровень Повышенный	На повышенном уровне: основными компьютерными программами и сетевыми ресурсами для обработки естественного языка, их функциональным инструментарием

3. СТРУКТУРА И СОДЕРЖАНИЕ КУРСА				
Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Объем в часах	Компетенции
	Раздел 1. Введение. История и современность компьютерной лингвистики	1/1	12	ОПК-5
1.1	<i>Компьютерная лингвистика или компьютер в лингвистике? /Лек/</i>	1/1	2	ОПК-5
1.2	Самостоятельная работа		10	ОПК-5
	Раздел 2. Цифровые текстовые корпусы и возможности их применения	1/1	15,95	ОПК-5
2.1	<i>Основные понятия корпусной лингвистики: введение. /Лек/</i>	1/1	1	ОПК-5
2.2	<i>Основные понятия корпусной лингвистики: инструменты поиска. /Лек/</i>	1/1	1	ОПК-5
2.3	<i>Применение корпусов для решения исследовательских задач. /Лек/</i>	1/1	2	ОПК-5
2.4	Самостоятельная работа	1/1	11,95	ОПК-5
	Раздел 3. Автоматизация перевода	1/1	4	ОПК-5
3.1	<i>Автоматический и автоматизированный перевод. /Лек/</i>	1/1	2	ОПК-5

3.2	<i>Системы переводческой памяти в работе и подготовке переводчиков. /Лек/</i>	1/1	2	ОПК-5
3.3	Самостоятельная работа	1/1	10	ОПК-5

4. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

4.1. Контрольные вопросы и задания

Раздел 1.

Контрольные вопросы.

1. С какими именами или открытиями связано становление компьютерной лингвистики как науки?

- a) Создание прототипа синтезатора речи в 1939 г.
- b) Ноам Хомский, профессор в Массачусетсе
- c) Цифровизация в гуманитарном образовании
- d) Роман Якобсон, профессор в Массачусетсе

2. Что такое NLP?

- a) Описание речи человека
- b) Цифровизация языка
- c) Исследование механизмов продуцирования естественного языка
- d) Обработка естественного языка

3. К какой области науки относится компьютерная лингвистика?

- a) Носит междисциплинарный характер
- b) Входит в группу компьютерных наук и информатики
- c) К гуманитарным наукам
- d) К естественным наукам

4. К какому направлению обучения в университетах России относится компьютерная лингвистика?

- a) Лингвистика
- b) Математика
- c) К любому из названных
- d) Такого направления обучения нет

Задания для самостоятельной работы

Сформулируйте 3 тезиса для выпускников школ «Почему я рекомендую вам поступить в университет для обучения по компьютерной / прикладной / математической лингвистике?»

Раздел 2.

1. Какие данные содержит метаразметка в текстовом корпусе?

- a) Грамматические признаки искомой словоформы
- b) Название документа, время его создания
- c) Номер страницы в документе, на которой размещен контекст
- d) Разметку синтаксической структуры контекста

1. В каком случае корпус является сбалансированным?

- a) В нем в равной степени представлена и устная, и письменная речь
- b) В нем есть и морфологическая, и синтаксическая разметка

- c) В нем представлен максимально широкий диапазон жанров письменной речи
- d) В нем представлены тексты различных жанров равного объёма в равной пропорции
2. В каком случае корпус называется поливариантным?
- a) В нем представлено несколько произведений одного автора
- b) В нем представлено произведение в оригинале и в нескольких переводах на один язык
- c) В нем представлено произведение в оригинале и переводах на разные языки
- d) В нем тот или иной язык представлен в нескольких вариантах
3. Какой функции пока нет в НКРЯ?
- a) Представление результатов в формате KWIC;
- b) Скачивание результатов в формате .docx;
- c) Сортировка результатов в случайном порядке;
- d) Сортировка результатов по дате создания.
4. Какие словоформы будут найдены по запросу лингвист* в окне «Поиск точных форм»?
- a) Все словоформы, в составе которых есть часть –лингвист-;
- b) Все словоформы, начинающиеся с лингвист-;
- c) Только все словоформы от леммы «лингвист»;
- d) Только контексты со словоформой «лингвист».
5. Какие словоформы будут найдены по запросу «любовь|ненависть» в окне «Лексико-грамматический поиск»?
- a) Запрос некорректный. Ничего не будет найдено;
- b) Контексты, в которых есть лемма «любовь», «ненависть» или обе эти леммы;
- c) Только контексты, в которых есть леммы «любовь» и «ненависть» в одном предложении;
- d) Только контексты, в которых есть словоформа «любовь» или «ненависть».
6. Как сформулировать запрос, чтобы найти все примеры с глаголом «радоваться» во всех формах в основном корпусе?
- a) «радоваться» в окне «Поиск точных форм»;
- b) «радоваться» в окне «Лексико-грамматический поиск: Слово»;
- c) «радов*» в окне «Поиск точных форм»;
- d) «радов*» в окне «Лексико-грамматический поиск: Слово».
7. Как найти первое употребление лексемы «счастье» в основном корпусе?
- a) Указать лексему в окне «Поиск точных форм», контексты упорядочить по дате создания;
- b) Указать лексему в окне «Лексико-грамматический поиск: Слово», контексты упорядочить по дате создания;
- c) Указать лексему в окне «Поиск точных форм», на странице с результатами выбрать пункт «формат KWIC»;
- d) Указать лексему в окне «Лексико-грамматический поиск: Слово», на странице с

результатами выбрать пункт «Статистика».

8. Как узнать, как переводчики переводили *достоевское* слово «надрыв» на английский язык?

a) Выбрать параллельный корпус (английский), в окне «Поиск точных форм» указать «надрыв»;

b) Выбрать параллельный корпус (английский), в окне «Лексико-грамматический поиск: Слово» указать «nadryw»;

c) Выбрать параллельный корпус (английский), в окне «Поиск точных форм» указать «надрыв|nadryw»;

d) Выбрать параллельный корпус (английский), в окне «Лексико-грамматический поиск: Слово» указать «надрыв».

Задания для самостоятельной работы

1. Найдите в Основном корпусе НКРЯ первый случай использования лексемы *Интернет*.

Опишите ход своей работы.

2. Найдите в Основном корпусе НКРЯ сочетание лемм *врать как...* во всех грамматических формах. Определите, какие лексемы чаще всего следуют за искомым словосочетанием. Опишите ход своей работы.

3. Определите на основе анализа Устного корпуса НКРЯ, какая форма произношения встречается чаще за последние десять лет: *Павлович* или *Палыч*. Опишите ход своей работы.

4. Определите на основе анализа многоязычного корпуса НКРЯ, как назван в переводах на русский язык произведения Л. Кэрролла «Алиса в стране чудес» персонаж, которого в оригинале зовут *Mock Turtle*. Опишите ход своей работы.

Раздел 3.

1. Каковы типичные ошибки систем машинного перевода?

2. Может ли машинный перевод использоваться в реальных ситуациях профессиональной коммуникации? Каковы условия этого?

3. В чем принципиальное отличие систем переводческой памяти от систем машинного перевода?

4. Каково обязательное условие использования систем переводческой памяти с точки зрения их обучаемости? Каким образом обеспечивается эта обучаемость?

5. С какой ключевой единицей текста работают системы переводческой памяти для обеспечения принципиальной возможности автоматизированного перевода?

6. Какие существуют возможности создания терминологической базы?

7. Какова основная форма взаимодействия выполняющих совместную переводческую работу (в терминах как социальной коммуникации, так и технологий компьютерной лингвистики (систем переводческой памяти))?

8. Каковы роли участников при выполнении общего переводческого заказа? Какие функции стоят за этими ролями?

9. Какие несоответствия выявляет автоматизированный контроль качества? В чем его назначение? Какой может быть возможная реакция на действия системы автоматизированного контроля качества?

10. Каковы главные условия сохранения формата текста оригинала в тексте перевода?
11. Чем может быть полезна концентрация внимания на отдельных единицах перевода, а не на всем тексте, при обучении переводу?

Задания для самостоятельной работы

Для выполнения заданий пользуйтесь демо-версией Memsource или связывайтесь с Высшей школой перевода Нижегородского государственного лингвистического университета для обеспечения возможности работы в академической версии с большим количеством функций.

1. Создайте пользователя/пользователей, при возможности – с разными ролями.
2. Создайте проект.
3. Загрузите файл, предварительно конвертировав его в MS Word.
4. При возможности (академическая версия) разбейте файл на несколько частей и распределите его между лингвистами (либо передайте одному).
5. Осуществите перевод, редакторскую правку перевода. При возможности (академическая лицензия) организуйте выполнение перевода командой и осуществите редакторскую правку всех переводов.
6. Осуществите автоматизированный контроль качества, отреагируйте соответствующим образом на выданные замечания.
7. Выгрузите готовый перевод как двуязычный файл и как готовый файл.

4.2. Фонд оценочных средств

Фонд оценочных средств представлен в Приложении 1

4.3. Перечень видов оценочных средств

Контрольные вопросы, задания для самостоятельной работы, автоматизированные тесты

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1. Рекомендуемая литература

5.1.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год
ЛП.1	<i>Бессмертный, И. А.</i>	Интеллектуальные системы : учебник и практикум для вузов / И. А. Бессмертный, А. Б. Нугуманова, А. В. Платонов. (Высшее образование). — ISBN 978-5-534-01042-8. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/511999	Москва : Издательство Юрайт, 2023. — 243 с.
ЛП.3	<i>Шунейко, А. А.</i>	Корпусная лингвистика : учебник для вузов / А. А. Шунейко. (Высшее образование). — ISBN 978-5-534-13603-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/519477	Москва : Издательство Юрайт, 2023. — 222 с.

5.1.2. Дополнительная литература

(литература из электронных ресурсов и печатные варианты из библиотеки НГЛУ)

	Авторы, составители	Заглавие	Издательство, год
ЛП.1	<i>Шунейко, А. А.</i>	Квантитативная лингвистика и новые информационные технологии : учебник для вузов / А. А. Шунейко, И. А. Авдеенко. (Высшее образование). — ISBN 978-5-534-15446-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/519734	Москва : Издательство Юрайт, 2023. — 347 с.

5.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

Э1	https://www.memsource.com/resources/
Э2	https://studiorum.ruscorpora.ru/

5.3. Перечень программного обеспечения

6.3.1.	Microsoft Windows 7, 10
6.3.2.	Microsoft Office - Word - Excel - Access - Power Point - Outlook
6.3.3.	Adobe Acrobat Reader DC

6.3.4.	ABBYY FineReader 11
6.3.5.	Microsoft Edge
6.3.6.	Mozilla Firefox
6.3.7.	Google Chrome
6.3.8.	CorelDraw
6.3.9.	Adobe InDesign CS6
6.3.10.	Adobe Photoshop
6.3.11.	Open Broadcaster Software (OBS Studio)
6.3.12.	Контур.Толк
6.3.13.	Система «Антиплагиат»
6.3.14.	1С:Бухгалтерия 8. Учебная версия
6.3.15.	Антивирус Касперского
5.4. Перечень информационных справочных систем	
6.4.1.	Справочная правовая система «Консультант Плюс»
6.4.2.	Справочная система «Гарант.Ру»

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ КУРСА

7.1	<i>Помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду вуза.</i>
-----	---

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ КУРСА

В дисциплине «Компьютерная лингвистика. Введение. Языковые корпусы. Автоматизированный перевод» (онлайн-курс) занятия требуют от студента интенсивной работы по освоению материала видеолекций лекций и вне аудитории, а именно:

- прослушивания лекций;
- самостоятельного решения практических задач;
- осуществления самоконтроля знаний;
- изучение рекомендуемой дополнительной литературы;
- взаимодействие с куратором (проктором) по решению технических и содержательных вопросов,
- выполнение автоматизированных тестов.