

Белорусский государственный университет



УТВЕРЖДАЮ

Проректор по учебной работе и
образовательным инновациям

О.Г. Прохоренко

1 декабря 2023 г.

Регистрационный № УД-797 /м.

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности**

7-06-0533-05 Прикладная математика и информатика

Профилизация: Интеллектуальные системы

2023 г.

Учебная программа составлена на основе ОСВО 7-06-0533-05-2023 (№ 160 от 18.05.23), примерного учебного плана 7-06-05-016/пр. от 18.01.2023 и учебного плана БГУ №М53-5.3-79/уч. от 11.04.2023 г.

СОСТАВИТЕЛИ:

Н.К. Рубашко – старший преподаватель кафедры информационных систем управления факультета прикладной математики и информатики Белорусского государственного университета

РЕЦЕНЗЕНТ:

Б.А. Железко – доцент кафедры маркетинга факультета маркетинга и предпринимательства БНТУ, кандидат технических наук

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой информационных систем управления Белорусского государственного университета (протокол № 3 от 19.10.2023 г.).

Научно-методическим Советом БГУ (протокол №3 от 30.11.2023 г.)

Заведующий кафедрой
информационных систем управления



В.В. Краснопрошин

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Учебная дисциплина «Компьютерная лингвистика» (КЛ) знакомит студентов магистратуры с теоретическими положениями лингвистического и математического моделирования естественного языка, с вариантами технологических решений, обеспечивающих компьютерную форму существования естественного языка. Дается лингвистическая интерпретация основных лингвистических объектов и единиц анализа, рассматриваются общие вопросы моделирования лингвистических сущностей. Дисциплина базируется на современных достижениях в области обработки естественного языка.

Цель учебной дисциплины – формирование у студентов магистратуры навыков системного представления о лингвистических основах информатики и аспектах моделирования языка и мышления в компьютерной среде.

Задачи учебной дисциплины:

1. создание теоретического базиса для использования современных технологий обработки данных при решении задач компьютерной лингвистики;

2. развитие у магистрантов навыков по проведению исследовательской работы, овладения ими методологией анализа результатов решения как теоретических, так и практических задач.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина относится к модулю «Основы искусственного интеллекта» компонента учреждения высшего образования.

Программа составлена с учетом **межпредметных связей** и программ по дисциплинам первой ступени высшего образования «Дискретная математика и математическая логика», «Методы и алгоритмы обработки данных», «Проектирование человеко-машинных интерфейсов» и дисциплины второй ступени высшего образования «Модели и методы искусственного интеллекта».

Требования к компетенциям

Освоение учебной дисциплины «Компьютерная лингвистика» должно обеспечить формирование следующих специализированных и углубленных профессиональных компетенций:

специализированные компетенции:

СК–2. Владеть методами и алгоритмами (в том числе интеллектуальными) решения задач поиска, распознавания и обработки данных;

СК–7. Уметь ставить и решать прикладные задачи компьютерной

лингвистики, определять методы и средства их эффективного решения;

СК–13. Уметь использовать научные и технические достижения для разработки эффективных алгоритмов решения прикладных задач;

Углубленные профессиональные компетенции:

УПК–5. Владеть перспективными технологиями программирования.

В результате изучения дисциплины студент магистратуры должен **знать:**

- основные понятия, составляющие базу современной лингвистической науки, и ее термины;
- методы и решения в системах организации знаний; эмпирические законы описания естественных языков;
- основные задачи и области применения компьютерной лингвистики;
- связь компьютерной лингвистики с другими дисциплинами прикладной математики и лингвистики;

уметь:

- пользоваться основными понятиями прикладной математики и лингвистики, применяемыми в области компьютерной лингвистики;
- реализовывать различные алгоритмы обработки естественного языка с использованием компьютера;

владеть:

- знаниями о системе языковедческих дисциплин; предмете, методах и моделях прикладной лингвистики; предмете компьютерной лингвистики;
- основными методами и приемами исследовательской и практической работы в области лингвистики.

Структура учебной дисциплины

Дисциплина изучается во 2 семестре. Всего на изучение учебной дисциплины «Компьютерная лингвистика» отведено:

- для очной формы получения высшего образования – 126 часов, в том числе 40 аудиторных часов, из них: лекции – 20 часов, лабораторные занятия – 20 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации по учебной дисциплине – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Краткое введение в проблематику КЛ

Тема 1.1. Введение в предмет КЛ

Объект исследования современной лингвистики текста. Информатика и компьютерная лингвистика. Область возникновения лингвистических проблем информатики. Основы теории речевой коммуникации. Совершенствование массовой и индивидуальной коммуникации. Особенности системы естественного языка (ЕЯ): уровни и связи.

Тема 1.2. Компьютерная лингвистика: методы, ресурсы, приложения

Характеризация классов задач компьютерной лингвистики. Приложения компьютерной лингвистики. WRE как обобщение регулярных выражений. Применение в компьютерной лингвистике

Раздел 2. Моделирование в компьютерной лингвистике

Тема 2.1. Моделирование языковых сущностей и человеческого мышления

Связь языка с мышлением. Элементы системы искусственного интеллекта. Модель механизма мышления. Ассоциативное построение понятий. Основные принципы работы системы и организации ее поведения.

Тема 2.2. Знание как объект моделирования

Представление знаний. Семантические сети и фреймы. Лингвистический аспект представления знаний. Понимание текста.

Тема 2.3. Естественный язык как объект моделирования

Особенности естественного языка как объекта моделирования. Классификация естественных языков.

Тема 2.4. Статистические закономерности языка

Основные статистические законы: закон Ципфа, закон Хипса, модель TF*IDF.

Раздел 3. Лингвистические ресурсы

Тема 3.1. Историческое развитие слова и его значений

Лексика современных текстов. Развитие терминологической лексики. Основной словарный состав. Выбор слов исходного массива. Выбор словарей.

Тема 3.2. Машиночитаемые словари

Традиционная и машинная лексикография. Отличия машинного словаря от обычного. Лемматизация. Составление машинных словарей. Контекстологический словарь как организатор базы знаний.

Тема 3.3. Корпус текстов и корпусная лингвистика

Основные понятия корпусной лингвистики. Основные характеристики корпусов. Проектирование и технологический процесс создания.

Стандартизация в корпусной лингвистике. Корпус как поисковая система. Обзор существующих корпусов различных типов. Корпусные исследования.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Очная форма получения углубленного высшего образования с применением
дистанционных образовательных технологий (ДОТ)

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов		Форма контроля знаний
		Лекции	Лабораторные занятия	
1	2	3	4	6
1	Краткое введение в проблематику КЛ	4	-	.
1.1	Введение в предмет КЛ	2	-	Устный опрос. Отчет по лабораторным работам
1.2	Компьютерная лингвистика: методы, ресурсы, приложения	2	-	
2	Моделирование в компьютерной лингвистике	6	8	
2.1	Моделирование языковых сущностей и человеческого мышления	2	2	Коллоквиум
2.2	Знание как объект моделирования	1	2	
2.3	Естественный язык как объект моделирования	1	2	Расчетно-графическая работа №1
2.4	Статистические закономерности языка	2	2	Расчетно-графическая работа №2
3	Лингвистические ресурсы	10	12	
3.1	Историческое развитие слова и его значений	2	-	Устный опрос
3.2	Машиночитаемые словари	4	6	Выполнение тестов. Расчетно-графическая работа №3
3.2	Корпус текстов и корпусная лингвистика	4	6	Отчет по лабораторным работам. Контрольная работа
	ВСЕГО:	20	20	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Волосатова Т. М. Информатика и лингвистика: Учебное пособие / Московский государственный технический университет им. Н.Э. Баумана. – Москва: ООО "Научно-издательский центр ИНФРА-М", 2023. – 196 с.
2. Прикладная и компьютерная лингвистика / под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. – Изд. 2-е. – Москва: URSS: Ленанд, 2017. – 315 с.
3. Потапова, Р. К. Новые информационные технологии и лингвистика учеб. пособие для студ. вузов, обуч. по спец. 021800 "Теоретическая и прикладная лингвистика" напр. 620200 "Лингвистика и новые информационные технологии" / Р. К. Потапова; Московский гос. лингвист. ун-т. - Изд. 6-е. - Москва: URSS: ЛЕНАНД, 2016. – 364 с.
4. Васильев, Ю. Обработка естественного языка. Python и SpaCy на практике / Юлий Васильев; [пер. с англ. И. Пальти]. – Санкт-Петербург; Москва; Минск: Питер, 2021. – 254 с. – <https://ibooks.ru/bookshelf/376835/reading>.

Перечень дополнительной литературы

1. Филиппович Ю.Н. Лингвистическое обеспечение информационных систем. Часть 1. Компьютерная лингвистика. Начало (посл. четв. XX века). Учебное пособие. М.: МГУП, 2013. – 452 с.
2. Автоматическая обработка текста на естественном языке и компьютерная лингвистика: учебное пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Песков О.В., Ягунова Е.В. – М.: МИЭМ, 2011.
3. Баранов А.Н. Введение в прикладную лингвистику. Изд.4, испр. и доп. –М.: УРСС, 2013. 3– 68 с.
4. Марчук Ю.Н. Компьютерная лингвистика. – М.: Изд-во АСТ. 2007 г. – 320 с.
5. Jurafsky, D. SpeechandLanguageProcessing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jurafsky, J. H. Martin. – New Jersey: Prentice Hall PTR, 2000. – 934 p.
6. Всеволодова А.В.Компьютерная обработка лингвистических данных. Изд.2 2007. 96 с.
7. Fastus: A cascaded finite-state transducer for extracting information from natural-language text / D. Israel [et al.] // Finite State Devices for Natural

- Language Processing / ed. by Roche, Schabes. – Cambridge, MA, USA: MIT Press, 1996. – P. 383–406.
8. CLAIM – научно-образовательный кластер. Серия книг "Компьютерная лингвистика". – Режим доступа:
<http://it-claim.ru/Library/Books/CL/CLbook.htm>
 9. Международная конференция Диалог. – Режим доступа: Dialog-21.ru
 10. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 22. 2023. С. 1–602. [Электронный ресурс]. – Режим доступа: <https://www.dialog-21.ru/media/5950/dialog2023scopuslr.pdf>
 11. Лаборатория компьютерной лингвистики Института проблем передачи информации РАН. – Режим доступа: Philol.msu.ru/~lex/library.htm

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенции в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: устный опрос, дискуссия.
2. Письменная форма: контрольные работы.
3. Устно-письменная форма: выполнение тестов, коллоквиум, расчетно-графические работы.

Формой текущей аттестации по дисциплине «Компьютерная лингвистика» учебным планом предусмотрен зачет.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку (формирование оценки за текущую успеваемость):

- расчетно-графические работы – 40 %;
- контрольные работы – 40 %;
- устный опрос, коллоквиум, тесты – 20%.

Отметка «зачет» выставляется магистранту, имеющему отметку за текущую успеваемость не ниже 4 («четырёх») баллов.

Примерная тематика лабораторных занятий

Занятие №1. Задачи компьютерной лингвистики

Занятие №2. Корпусная лингвистика. Особенности представления.

Занятия №3. Статистические закономерности естественного языка.

Занятие №4. Создание частотного индекса на заданном корпусе текстов.

Занятие №5. Разработка лексической онтологии заданного корпуса текстов.

Занятие №6. Корпус как поисковая система.

Занятие №7. Исследования с использованием Национального корпуса русского языка.

Занятие №8. Классификаторы естественного языка.

Занятие №9. Технологии создания машинных словарей.

Занятие №10. Понятие лингвистического процессора.

Рекомендуемая тематика расчетно-графических работ

Работа 1. Разработка частотных словарей для любых трех языков.

Работа 2. Исследование статистических закономерностей языка.

Работа 3. Статистический анализ текстов из Национального корпуса русского языка.

Рекомендуемая тематика контрольной работы

Контрольная работа. Технологии создания машинных словарей.

Текущий контроль знаний проводится в соответствии с учебно-методической картой дисциплины.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса большинства практических занятий используется *практико-ориентированный подход*, который предполагает освоение содержания учебного материала через решение практических задач, а также приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Кроме этого, при организации образовательного процесса используется комбинация таких методов *креативного обучения*, как *методы группового обучения, проектного обучения и учебной дискуссии*. Комбинация методов предполагает ориентацию на генерирование идей, приобретение навыков для решения исследовательских, творческих и коммуникационных задач, появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

Методические рекомендации по организации самостоятельной работы обучающихся, подготовка к экзамену

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине используется образовательный портал БГУ <https://edufpmi.bsu.by>, на котором размещаются комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебные издания для теоретического изучения дисциплины, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к экзамену, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Примерный перечень вопросов к зачету

1. Предмет компьютерной лингвистики.
2. История становления и основные тенденции развития.
3. Основные направления компьютерной лингвистики.
4. Компьютерное обеспечение представления знаний.
5. Автоматизированный анализ: распознавание и синтез устной и письменной речи.
6. Морфологический анализ, проблемы семантического анализа, синтаксический анализ.
7. Лингвистические базы данных: модели и типы данных.
8. Создание общих искусственных языков для представления информации.
9. Компьютерная лексикография как одно из направлений прикладной лингвистики.
10. Словарные процессоры.
11. Основные понятия структуры словаря: словник, словарная статья, грамматические, стилистические пометы; иллюстративный материал.
12. Типология электронных словарей.
13. Тезаурусы и терминологические словари.
14. Компьютерные технологии составления и эксплуатации словарей.
15. Корпусная лингвистика, ориентированная на создание национальных корпусов естественных языков.
16. Типы лингвистической разметки.
17. Концепция обучения на основе компьютерных технологий.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Технологии обработки текста	Информационных систем управления	Нет	Оставить содержание учебной дисциплины без изменения, (протокол № 3 от 19 октября 2023 г.)

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на ____ / ____ учебный год

№№ Пп	Дополнения и изменения	Основание

Заведующий кафедрой

(степень, звание)

(подпись)

(И.О.Фамилия)

УТВЕРЖДАЮ
Декан факультета

(степень, звание)

(подпись)

(И.О.Фамилия)