БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ

Ректор Белорусского
государственного университета

_____А.Д.Король

27 июня 2025 г.

Регистрационный № 3613/б.

ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА

Учебная программа учреждения образования по учебной дисциплине
для специальности:

6-05-0533-11 Прикладная информатика

Профилизация: Информационные аналитические системы

2025 г.

Учебная программа составлена на основе ОСВО 6-05-0533-11-2023, учебного плана БГУ № 6-5.3-59/03 от 15.05.2023

СОСТАВИТЕЛЬ:

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

РЕЦЕНЗЕНТЫ:

С.В.Абламейко, профессор кафедры веб-технологий и компьютерного моделирования механико-математического факультета Белорусского государственного университета, академик НАН Беларуси, доктор технических наук, профессор;

В.А.Головко, заведующий кафедрой интеллектуальных информационных технологий учреждения образования «Брестский государственный технический университет», доктор технических наук, профессор

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой информационных систем управления БГУ
(протокол № 15 от 19.06.2025);

Научно-методическим советом БГУ
(протокол № 11 от 26.06.2025)

Заведующий кафедрой _____ А.М.Недзьведь

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины – формирование у студентов навыков системного представления об основах вычислительной (компьютерной) лингвистики и аспектах моделирования языка и мышления в компьютерной среде, формирование базы, необходимой для дальнейшего самостоятельного углубленного изучения вычислительной лингвистики.

Учебная дисциплина профилизации «Вычислительная лингвистика» (ВЛ) знакомит студентов с основными направлениями вычислительной (компьютерной) лингвистики, подходами к формированию корпусов текстов, разработке лингвистических баз знаний, применением инструментария вычислительной лингвистики к автоматизации лингвистических исследований. В программу дисциплины включены как классические подходы к компьютерному анализу текста, так и ряд методов, разработанных автором данного курса.

Задачи учебной дисциплины:

1. создание теоретического базиса для использования современных технологий обработки информации при решении задач вычислительной (компьютерной) лингвистики;
2. развитие у студентов навыков по проведению исследовательской работы, овладение ими методологии решения как теоретических, так и практических задач и анализа результатов.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием.

Учебная дисциплина относится к дисциплинам профилизации «Информационные аналитические системы» компонента учреждения образования.

Связи с другими учебными дисциплинами, включая учебные дисциплины компонента учреждения высшего образования, дисциплины специализации и др.

Программа составлена с учетом межпредметных связейи программ по учебным дисциплинам первой ступени высшего образования «Дискретная математика и математическая логика», «Методы и алгоритмы обработки данных», «Проектирование человеко-машинных интерфейсов», «Нейросетевые технологии обработки данных».

Требования к компетенциям

Освоение учебной дисциплины «Вычислительная лингвистика» должно обеспечить формирование следующих компетенций*:*

Универсальные компетенции:

Решать стандартные задачи профессиональной деятельности на основе применения информационно-коммуникационных технологий.

Специализированные компетенции:

Ставить и решать прикладные задачи компьютерной лингвистики, определять методы и средства их эффективного решения.

В результате освоения учебной дисциплины студент должен:

знать:

– основные понятия, составляющие базу современной лингвистической науки, и ее термины;

– основные задачи, решаемые компьютерной лингвистикой;

– базовые подходы к автоматизации лингвистических исследований в компьютерной лингвистике;

– связь компьютерной лингвистики с другими дисциплинами прикладной информатики и лингвистики;

уметь:

– пользоваться основными понятиями прикладной информатики и лингвистики, применяемыми в области компьютерной лингвистики;

– реализовывать различные алгоритмы обработки естественного языка с использованием компьютера;

иметь навык:

– применения основных методов и приемов исследовательской и практической работы в области компьютерной лингвистики;

– эффективного решения прикладных задач компьютерной лингвистики;

– самостоятельного построения как систем автоматизации лингвистических исследований, так и прикладных систем обработки естественного языка.

Структура учебной дисциплины

Дисциплина изучается в 5 семестре. В соответствии с учебным планом всего на изучение учебной дисциплины «Вычислительная лингвистика» отведено **для очной формы** получения высшего образования – 108 часов, в том числе 68 аудиторных часов, лекции –34 часа, лабораторные занятия –34 часа.

Из них:

Лекции – 34 часа, лабораторные занятия – 30 часов, управляемая самостоятельная работа (УСР) – 4 часа.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации– экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1 Краткое введение в проблематику ВЛ

Тема 1.1 Введение в предмет ВЛ

Объект исследования современной лингвистики текста. Информатика и вычислительная лингвистика (ВЛ). Область возникновения лингвистических проблем информатики. Основные направления вычислительная лингвистики. Компьютерный анализ текста.

Тема 1.2 Особенности естественного языка как объекта моделирования

Естественный язык (ЕЯ) как универсальное средство описания действительности и коммуникации с вычислительной системой. Характеристика естественного языка как объекта моделирования и его особенности. Классификация естественных языков.

Тема 1.3 Лингвистические ресурсы ВЛ

Виды ресурсов, информационные ресурсы и их электронное представление. Понятие и виды лингвистических ресурсов. Лингвистические ресурсы как знания о ЕЯ. Лингвистические базы знаний (ЛБЗ): состав, технологии.

Раздел 2 Инструментарий вычислительной лингвистики

Тема 2.1 Классификаторы свойств ЕЯ

Классификаторы как метод уменьшения сложность задачи путем преобразования нечетких лингвистических объектов в дискретные лингвистические единицы. Понятие классификации и классификатора. Методы классификации, особенности, их преимущества и недостатки. Принципы разработки классификаторов ЕЯ в зависимости от их типологии.

Тема 2.2 Словари естественного языка

Типы словарей ЕЯ: алфавитный, частотный, обратный, синонимов, антонимов и др. Назначение, состав, правила составления словарей ЕЯ. Понятие машинного словаря. Отличия машинного словаря от обычного. Лемматизация. Составление машинных словарей. Структура, назначение, использование в современных системах обработки ЕЯ.

Тема 2.3 Корпусная лингвистика как составная часть вычислительной лингвистики

Основные понятия корпусной лингвистики. Текст как основной источник знаний. Понятие корпуса текстов, его основные характеристики. Корпус текстов как особый лингвистический ресурс. Исходный и аннотированный корпус текстов. Назначение, принципы составления. Проектирование и технологический процесс создания.

Стандартизация в корпусной лингвистике. Корпус как поисковая система. Корпусные исследования.

Тема 2.4 Статистические закономерности ЕЯ

Основные аспекты статистических закономерностей естественного языка. Характеристики статистических законов языка: закона Ципфа, закона Хипса. Применение статистических методов для анализа языка.

Тема 2.5 Понятие базового лингвистического процессора

Принципы построения базового лингвистического процессора (ЛП). Тестирование. Общепринятая схема лингвистического анализа текста. Преформатор текстовых документов. Лексический анализ текстовых документов. Лексико-грамматический, синтаксический и семантический анализ текста.

Раздел 3 Задачи лингвистических информационных технологий

Тема 3.1 Машинный перевод текста

Общая постановка задачи машинного перевода (МП). Основные стратегии ее решения. Анализ и синтез текста в задаче МП. Оптимизация ЛБЗ системы МП в условиях многоязычной информационной среды. Понятие interlingua и возможные способы ее реализации. МП в условиях близости входного и выходного языков.

Тема 3.2 Информационный поиск

Структурно-функциональная схема системы информационного поиска. Автоматизация индексирования текстовых документов и запросов пользователя. ЕЯ-интерфейс пользователя. Оценка эффективности систем информационного поиска, оптимизация алгоритмов.

Тема 3.3 Автоматическое реферирование текста

Общая постановка задачи автоматического реферирования текста (АРТ). Основные типы рефератов: классический, topic-ориентированный, запросно-ориентированный. Алгоритмы автоматического построения основных типов рефератов.

Раздел 4 Нейронные сети в лингвистике

Тема 4.1 Краткое введение в нейронные сети

Обзор нейронных сетей и их применения. Краткая история нейронных сетей в лингвистическом анализе. Основные понятия и терминология. Применение нейронных сетей в лингвистическом анализе. Математические основы нейронных сетей.

Тема 4.2 Векторная модель представления текстовой информации

Определение векторного представления. Обзор методов векторного представления текстов: метод набора слов, TF, TF-IDF. Проблемы разреженных векторных моделей. Векторное пространство. Вектор тем документа.

Тема 4.3 Применение машинного обучения для обработки естественного языка

Особенности машинного обучения для ЕЯ. Подходы к машинному обучению. Машинное обучение с учителем. Машинное обучение без учителя.

Создание обучающих примеров данных. Характеристика моделей Word2Vec, GloVe (модель глобальных векторов), fastText.

Тема 4.4 Классификация текстов

Описание методов классификации. Оценка качества классификации. Вероятностные методы классификации. Методы классификации на основе расстояний. Методы классификации на основе правил. Комбинированные методы классификации. Метод комбинированной иерархической классификации.

Тема 4.5 Кластерный анализ текста

Общее описание методов кластерного анализа. Оценка качества кластерного анализа. Вероятностные методы кластерного анализа. Структурные методы кластерного анализа. Интерпретация результатов кластерного анализа. Сравнительный анализ методов кластерного анализа.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

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

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
1	**Краткое введение в проблематику ВЛ**	**6**			**2**			
1.1	Введение в предмет ВЛ	2						
1.2	Особенности естественного языка как объекта моделирования	2						Устный опрос
1.3	Лингвистические ресурсы ВЛ	2			2			Лабораторная работа
2	**Инструментарий вычислительной лингвистики**	**12**			**12**		**2**	
2.1	Классификаторы свойств ЕЯ	2			2			Лабораторная работа
2.2	Словари естественного языка	2			4			Расчетно-графическая работа
2.3	Корпусная лингвистика как составная часть вычислительной лингвистики	2					2	Промежуточный тест на портале
2.4	Статистические закономерности ЕЯ	2			4			Расчетно-графическая работа

2.5	Понятие базового лингвистического процессора	4			2			Лабораторная работа
3	**Задачи лингвистических информационных технологий**	**6**			**4**		**2**	
3.1	Машинный перевод текста	2			2			Лабораторная работа
3.2	Информационный поиск	2			2			Расчетно-графическая работа
3.3	Автоматическое реферирование текста	2					2	Промежуточный тест на портале
4	**Нейронные сети в лингвистике**	**10**			**12**			
4.1	Краткое введение в нейронные сети	2						Устный опрос
4.2	Векторная модель представления текстовой информации	2			4			Расчетно-графическая работа
4.3	Применение машинного обучения для обработки ЕЯ	2			4			Лабораторная работа
4.4	Классификация текстов	2			2			Лабораторная работа
4.5	Кластерный анализ текста	2			2			Итоговый тест на образовательном портале
	ВСЕГО:	**34**			**30**		**4**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Основная литература

1. Волосатова Т. М. Информатика и лингвистика: Учебное пособие / Московский государственный технический университет им. Н.Э. Баумана. – Москва: ООО "Научно-издательский центр ИНФРА-М", 2025. – 196 с. – URL: https://znanium.com/catalog/document?id=422587.

2. 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. –URL:https://ibooks.ru/bookshelf/376835/reading.

Дополнительная литература

1. Хобсон, Лейн. Обработка естественного языка в действии = NaturalLanguageProcessinginAction / Л. Хобсон, Х. Ханнес, Х. Коул; [пер. с англ. И. Пальти, С. Черников]. – Санкт-Петербург: Питер, 2021. – 576 с. – [Электронный ресурс]. – Режим доступа: https://ibooks.ru/bookshelf/371695.

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. Филиппович Ю.Н. Лингвистическое обеспечение информационных систем. Часть 1. Компьютерная лингвистика. Начало (посл. четв. XX века). Учебное пособие. М.: МГУП, 2013. – 452 с.

5. Автоматическая обработка текста на естественном языке и компьютерная лингвистика: учебное пособие / Большакова Е.И., Клышинский Э.С., ЛандэД.В. , Носков А.А., Песков О.В., Ягунова Е.В. – М.: МИЭМ, 2011.

6. Баранов А.Н. Введение в прикладную лингвистику. Изд.4, испр. и доп. –М.: УРСС, 2013. 3– 68 с.

7. Марчук Ю.Н. Компьютерная лингвистика. – М.: Изд-во АСТ. 2007 г. – 320 с.

8. CLAIM – научно-образовательный кластер. Серия книг "Компьютерная лингвистика". – Режим доступа: http://it-claim.ru/Library/Books/CL/CLbook.htm

9. Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной международной конференции «INTERNATIONAL CONFERENCE onComputationalLinguisticsandIntellectualTechnologies». [Элекстронный ресурс]. – Режим доступа: https://dialogue-conf.org/

10. Лаборатория компьютерной лингвистики Института проблем передачи информации РАН. – Режим доступа: Philol.msu.ru/~lex/library.htm

Перечень рекомендуемых средств диагностики и методика формирования итоговой отметки

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

1. Устная форма: устный опрос.
2. Письменная форма: расчетно-графические работы, лабораторные работы.
3. Устно-письменная форма: выполнение тестов на образовательном портале БГУ.

Формой промежуточной аттестации по дисциплине «Вычислительная лингвистика» учебным планом предусмотрен экзамен.

Для формирования итоговой отметки по учебной дисциплине используется модульно-рейтинговая система оценки знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая система предусматривает использование весовых коэффициентов для текущей и промежуточной аттестации студентов по учебной дисциплине.

Формирование итоговой отметки в ходе проведения контрольных мероприятий текущей аттестации (примерные весовые коэффициенты, определяющие вклад текущей аттестации в отметку при прохождении промежуточной аттестации):

- расчетно-графические работы – 60 %;
- выполнение теста – 40%.

Итоговая отметка по дисциплине рассчитывается на основе итоговой отметки текущей аттестации (модульно-рейтинговой системы оценки знаний) 40 % и экзаменационной отметки 60 %.

Примерный перечень заданий для управляемой самостоятельной работы

Тема 2.3 Корпусная лингвистика как составная часть вычислительной лингвистики (2 ч)

Создание и разметка (аннотирование) корпусов текстов и разработка средств поиска по ним, экспериментальные исследования на базе корпусов.

(Форма контроля – выполнение промежуточного теста на образовательном портале).

Тема 3.3 Автоматическое реферирование текста (2 ч)

Методы автоматического реферирования текстов и их практическая реализация.

(Форма контроля – выполнение промежуточного теста на образовательном портале).

Примерный перечень лабораторных занятий

1. Задачи вычислительной (компьютерной) лингвистики.
2. Статистические закономерности естественного языка (УСР).
3. Создание частотного индекса на заданном корпусе текстов.
4. Исследования с использованием Национального корпуса русского языка.
5. Классификаторы естественного языка.
6. Технологии создания машинных словарей.
7. Лингвистический процессор и его функциональность.
8. Методы машинного перевода.
9. Исследование популярных ИПС, изучение расширенной функциональности для поиска документов и веб-страниц.
10. Нейросетевая обработка текста.
11. Векторное представление текста.
12-13. Задачи, решаемые с помощью векторного представления текста.
14. Методы классификации.
15. Методы кластеризации.

Рекомендуемая тематика расчетно-графических работ

1. Разработка частотных словарей для любых трех языков.
2. Исследование статистических закономерностей языка.
3. Векторная модель текста.
4. Исследование модели Word2Vec.

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

Для организации самостоятельной работы студентов 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 лабораторным занятиям, электронные версии заданий для расчетно-графических работ, электронные тесты, 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, список рекомендуемой литературы, информационных ресурсов и др.).

Примерный перечень вопросов к экзамену

1. Компьютерная лингвистика как научное направление
2. Особенности ЕЯ как объекта моделирования
3. Понятие корпусной лингвистики
4. Понятие лингвистической базы знаний. Краткая характеристика
5. Понятие исходного корпуса текстов. Принципы формирования
6. Типы словарей ЕЯ
7. Классификаторы лингвистической базы знаний. Особенности разработки
8. Разработка классификатора для флективных языков
9. Понятие базового словаря
10. Корпус аннотированных текстов. Процесс порождения
11. Корпус параллельных текстов. Особенности
12. Базовый лингвистический процессор. Состав и принципы построения
13. Этапы лингвистического анализа текста
14. Преформатор текстовых документов
15. Лексический анализ текстовых документов
16. Лексико-грамматический анализ текстов
17. Принципы синтаксического анализа текстов
18. Характеристика семантического анализа текстов
19. Задача информационного поиска. Основные понятия
20. ИПС: определение, общая схема функционирования, основные типы
21. Классификация ИПЯ. Главные компоненты ИПЯ
22. Особенности поиска по ключевым словам
23. Особенности поиска по естественно-языковому запросу
24. Особенности поиска по заданному документу
25. Статистический метод автоматического реферирования на основе САО
26. Общая характеристика задачи машинного перевода
27. Метод прямого машинного перевода
28. Трансферный метод машинного перевода текстов
29. Методы интерлигвы в машинном переводе

30. Проблема многозначности при машинном переводе
31. Векторная модель текста. Особенности
32. Понятие векторного пространства
33. Метод "мешок слов" векторного представления текста
34. Функции взвешивания в векторных моделях текста
35. Тематическое моделирование. Модель TF*IDF
36. Применение нейронных сетей в автоматической обработке текстов
37. Подходы к машинному обучению в обработке ЕЯ
38. Машинное обучение с учителем
39. Машинное обучение без учителя
40. Общая характеристика модели Word2Vec
41. Особенности архитектуры CBoW модели Word2Vec
42. Особенности архитектуры Skip-gram модели Word2Vec
43. Особенности модели GloVe
44. Особенности модели fastText
45. Классификация документов. Основные алгоритмы
46. Кластеризация. Алгоритмы

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УО

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)
Теория распознавания образов	Кафедра информационных систем управления	Предложения отсутствуют	Рекомендовать к утверждению учебную программу (протокол № 15 от 19.06.2025)
Нейросетевые технологии обработки данных	Кафедра информационных систем управления	Предложения отсутствуют	Рекомендовать к утверждению учебную программу (протокол № 15 от 19.06.2025)

Заведующий кафедрой
информационных систем управления, _____ А.М.Недзьведь
доктор технических наук, доцент

19.06.2025

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ УО

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 202_ г.)

Заведующий кафедрой
_____ _____

УТВЕРЖДАЮ
Декан факультета
_____ _____