

Министерство науки и высшего образования
Российской Федерации

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Донецкий государственный университет»

Филологический факультет

Кафедра общего, славянского и прикладного языкознания имени Е.С. Отина

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат 03C1BBABF83E153235592ADEEC463FDD
Владелец МАШАРОВ ПАВЕЛ АНАТОЛЬЕВИЧ
Действителен с 06.07.2026 по 29.09.2027
Должность проректор

УТВЕРЖДАЮ
проректор

«17» апреля 2025 г.
МП П.А. Машаров

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ЕСТЕСТВЕННЫХ ЯЗЫКОВ

Углубленная группа направлений подготовки	45.00.00 Языкознание и литературоведение
Программа высшего образования	Программа бакалавриата
Направление подготовки	45.03.03 Фундаментальная и прикладная лингвистика
Направленность образовательной программы (профиль)	Фундаментальная и прикладная лингвистика
Квалификация	Бакалавр
Форма обучения	Очная

Рабочая программа может быть адаптирована для лиц
с ограниченными возможностями здоровья и инвалидов

Донецк 2025

Рабочая программа дисциплины **«Автоматическая обработка естественных языков»** для обучающихся по направлению подготовки 45.03.03 Фундаментальная и прикладная лингвистика (Профиль: Фундаментальная и прикладная лингвистика), составлена на основании Федерального государственного образовательного стандарта высшего образования – бакалавриат по направлению подготовки 45.03.03 Фундаментальная и прикладная лингвистика, утвержденного приказом Министерства образования и науки Российской Федерации от 24 апреля 2018 г. № 323 (с изм. и доп.), Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры, утвержденного приказом Министерства науки и высшего образования Российской Федерации от 06 апреля 2021 г. № 245 (с изм. и доп.), в соответствии с учебным планом, утвержденным Ученым советом ФГБОУ ВО «ДонГУ» для набора 2025 года.

Разработчики:

доцент кафедры общего, славянского и
прикладного языкознания имени Е. С. Отина,
канд. филол. наук

Э. А. Титаренко

ассистент кафедры общего, славянского и
прикладного языкознания имени Е. С. Отина

Д. А. Борозенец

Рабочая программа одобрена на заседании кафедры общего,
славянского и прикладного языкознания имени Е. С. Отина,
Протокол от 03.04.2025 г. № 8.
Заведующий кафедрой

Н. А. Ярошенко

СОГЛАСОВАНО:

Декан филологического факультета
15.04.2025 г.

Н. А. Ярошенко

Учебно-методическая комиссия филологического факультета.
Протокол от 15.04.2025 г. № 4.
Председатель

А. Н. Стебунова

Руководитель основной
образовательной программы,
канд. филол. наук, доцент
15.04.2025 г.

Н. А. Ярошенко

1. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

1.1. Требования к предварительной подготовке обучающихся, предшествующие и сопутствующие дисциплины, на которых основывается изучение данной:

дисциплины программы бакалавриата: Основы прикладной лингвистики, Информационные технологии в лингвистике, Компьютерная лингвистика.

1.2. Дисциплины, курсовые работы и практики, для которых освоение данной дисциплины необходимо как предшествующее: Учебная практика: проектно-технологическая практика, Квантитативная лингвистика.

2. ОПИСАНИЕ ДИСЦИПЛИНЫ

2.1. Общая характеристика

Наименование показателя	Значение показателя
Название образовательной программы	45.03.03 Фундаментальная и прикладная лингвистика (Профиль: Фундаментальная и прикладная лингвистика)
Шифр и название в соответствии с учебным планом	Б1.Б.М7 Автоматическая обработка естественных языков
Часть образовательной программы	Базовая часть
Количество зачетных единиц / всего часов	6 / 216

В случае предъявления от обучающегося или его родителя (законного представителя) заявления на обучение по адаптированной образовательной программе высшего образования, подкрепленного заключением психолого-медико-педагогической комиссии (ПМПК) или медико-социальной экспертизы (МСЭ) с рекомендациями создания индивидуальной программы реабилитации и абилитации (ИПРА), данная рабочая программа может быть адаптирована с учетом индивидуальных особенностей здоровья обучающегося.

2.2. Распределение часов по формам и периодам обучения

Форма обучения	курс	семестр	Общее количество часов					Форма контроля
			лекционных	лабораторных	практических	самостоятельной работы + контроль	всего	
Очная	3	5	34	51	–	59	144	зачет
Очная	3	6			48	24	72	экзамен
Очная, всего			34	51	48	83	216	

3. ЦЕЛИ ДИСЦИПЛИНЫ

Данная дисциплина имеет цель подготовить студентов к эффективному использованию компьютерных средств для решения задач в профессиональной деятельности, а также совершенствование подготовки студентов в области аналитической работы с использованием документальных информационных ресурсов.

**4. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ
ОСВОЕНИЯ КОМПОНЕНТА ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ, ИХ ИНДИКАТОРЫ
И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ**

Компетенции	Индикаторы	Результаты обучения
ОПК-2. Способен к ведению профессиональной деятельности с опорой на основы математических дисциплин, необходимых для формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур	ОПК-2.2. Соотносит задачи формализации лингвистических знаний с основами математических знаний.	ОПК-2.2.1. Знает основные уровни обработки текста на ЕЯ и существующие модели статистического, морфологического и синтаксического анализа текстов. ОПК-2.2.2. Умеет пользоваться лингвистически ориентированными программными продуктами. ОПК-2.2.3. Владеет навыками работы со специальными программными средствами автоматизированной обработки текстов. - современными методами и средствами обработки и хранения информации; - навыками работы с компьютером как средством управления информацией; - способностью работать с информацией в глобальных компьютерных сетях.

5. ПРОГРАММА ДИСЦИПЛИНЫ

Название темы	Краткое содержание темы (вопросы темы)
Раздел 1.	
Тема 1. Введение в обработку естественно-языковых текстов. Разновидности языковых групп и особенности их обработки.	Представление об уровнях представления языка – фонетика, морфология, синтаксис, семантика. Лингвистика и прагматика. Лингвистическое моделирование. Действующие модели языка. Теория «Смысл –Текст» как фундамент для построения систем автоматической обработки текста. Понятие токенизации. Основные подходы. Автоматические разбиение текста на предложения и другие структурные блоки. Основные проблемы и методы. Особенности сегментации русскоязычных текстов.
Тема 2. Методы обработки естественных языков.	Анализ и синтез текста. Представление знаний для компьютерной обработки. Методы классификации и кластеризации
Тема 3. Задачи автоматического анализа текста.	Прикладные задачи автоматической обработки текстов. Задача определения авторства текстов. Задача определения тематической принадлежности текстов.

	Задача анализа тональности текстов.
Тема 4. Лексические ресурсы, применяемые NLP-системами:	Словари, тезаурусы, онтологии
Тема 5. Классическая схема анализа текста: графематический анализ, морфологический анализ, синтаксический анализ, семантический анализ.	Морфологический и синтаксический анализ. Парсинг. Морфологический разбор, стемминг, лемматизация. Понятия лексемы, словоформы, леммы, морфемы, псевдо-основы и псевдо-окончания. Грамматические категории. Словоизменительная парадигма. Грамматические модели русского языка в контексте автоматической обработки. Различные подходы к синтаксическому анализу: анализ «сверху вниз» и «снизу вверх». Алгоритм синтаксического анализа. Синтаксические отношения. Деревья зависимостей.
Тема 6. Категоризация способов использования NLP	Поиск, веб-поиск, поиск по документам, автозаполнение.
Тема 7. Аннотирование и реферирование текста	Подходы и методы автоматического реферирования: статистические методы метод Освальда методы статистических ассоциаций. логико-математические методы метод текстовых связей. Автоматическое обобщение MS Word 2003, 2007. ИИ-инструменты для резюмирования документов
Тема 8. Машинный перевод	Основные формы взаимодействия человека и компьютера при машинном переводе. Анализ наиболее распространенных систем машинного перевода
Тема 9. Основные понятия корпусной лингвистики	Национальный корпус текстов. Типы корпусов. Понятие сбалансированного vs. мониторного корпуса. Требования к корпусам: репрезентативность, полнота, структурированность. Единицы хранения. Типы аннотации корпусов. Назначение корпусов. Использование корпусов при разработке систем автоматического анализа текстов
Раздел 2	
Тема 7. Автоматическая обработка естественного языка в научно-исследовательской деятельности.	Исследовательские лингвистические информационные технологии. Корпусно-ориентированный подход в лингвистике. Использование корпусов текстов в научных исследованиях. Типы лингвистических информационных технологий и их взаимодействие
Тема 8. Автоматическая обработка естественного языка в образовательной деятельности.	Вопросно-ответные системы. Извлечение информации из текста: извлечение именованных сущностей, отношений и фактов Автоматическое создание контента. Автоматизированные системы оценки эссе.

6. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

6.1.Форма обучения – очная, курс – 3, семестр – 5

Наименования разделов и тем	Количество часов				
	Лекц.	Лабор.	Практ.	СРС+К	Всего
Тема 1. Введение в обработку естественно-языковых текстов. Разновидности языковых групп и особенности их обработки.	4	6		14	24
Тема 2. Методы обработки естественных языков.	6	10		8	24
Тема 3. Задачи автоматического анализа текста.	6	10		8	24
Тема 4. Лексические ресурсы, применяемые NLP-системами:					
Тема 5. Классическая схема анализа текста: графематический анализ, морфологический анализ, синтаксический анализ, семантический анализ.					
Тема 6. Категоризация способов использования NLP					
Тема 7. Аннотирование и реферирование текста	6	8		10	24
Тема 8. Машинный перевод	6	8		10	24
Тема 9. Основные понятия корпусной лингвистики	6	9		9	24
ИТОГО ЗА СЕМЕСТР	34	51	–	93	144

6.1.Форма обучения – очная, курс – 3, семестр – 6

Наименования разделов и тем	Количество часов				
	Лекц.	Лабор.	Практ.	СРС+К	Всего
Тема 7. Автоматическая обработка естественного языка в научно-исследовательской деятельности.		24		12	36
Тема 8. Автоматическая обработка естественного языка в образовательной деятельности.		24		12	36
ИТОГО ЗА СЕМЕСТР	0	48	–	93	72
ИТОГО ПО КОМПОНЕНТУ ОП	34	51	48	93	144

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ (СРЕДСТВА) ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

7.1. Контрольные вопросы

Раздел 1

Трудности автоматизации обработки естественного языка в интеллектуальных системах.

Этапы анализа предложений на естественном языке.

Синтаксически-ориентированный и семантически-ориентированный анализ ЕЯ.

Основные методы анализа ЕЯ: шаблоны.

Основные методы анализа ЕЯ: семантические грамматики.

Основные методы анализа ЕЯ: падежные фреймы.

Деревья анализа и свободно-контекстные грамматики.

Задачи автоматического анализа текста. Проверка правописания, грамматики и стиля. Распознавание текстов. Распознавание и синтез речи. Машинный перевод текста и речи (классика NLP). Информационный поиск. Реферирование. Классификация (кластеризация и рубрикация) текстов, установление сходства текстов (плагиат и т.п.). Автофильтрация. Вопросно-ответные и диалоговые системы. Системы извлечения знаний (извлечение именованных сущностей, извлечение отношений и фактов, извлечение тональности/мнений). Примеры систем

Типы лингвистических данных. Типы электронных ресурсов, представляющих лингвистические данные: лексикографические ресурсы; корпуса. Примеры.

Психолингвистический подход к анализу естественного языка.

Раздел 2

Корпуса текстов. Основные понятия корпусной лингвистики. Корпус текстов: определение, решаемые задачи. Национальный корпус текстов. Типы корпусов. Понятие сбалансированного vs. мониторингового корпуса. Требования к корпусам: репрезентативность, полнота, структурированность. Единицы хранения. Типы аннотации корпусов. Назначение корпусов. Использование корпусов при разработке систем автоматического анализа текстов.

Этапы предварительной обработки текста. Сбор корпуса текстов. Проблема унификации текстов в корпусе (кодировка, «многозначность» служебных символов, нетекстовые элементы и т.п.). Стандарты представления текстов (например, TEI). Предварительная обработка текста. Задачи и проблемы токенизации. Особенности токенизации для разных типов и жанров текстов. Сегментация текстов на предложения. Методы сегментации.

7.2. Темы письменных работ (типы заданий)

История развития компьютерной лингвистики и АОЯ. Анализ ключевых этапов: от символических методов до современных нейронных сетей.

Уровни анализа в АОЯ. Разбор задач на разных ступенях: графематический анализ, токенизация, морфологический анализ (стемминг, лемматизация), синтаксический и семантический анализ.

Сравнение подходов к АОЯ. Сопоставление символического, статистического и нейросетевого подходов с точки зрения их сильных и слабых сторон.

Извлечение информации и построение онтологий. Технологии для извлечения структурированных данных из текстов.

Контрольная работа по проверке теоретических знаний – по всем темам, с использованием указанных выше контрольных вопросов.

7.3. Образец содержания экзаменационного билета 6 семестр

1. Методы прикладной лингвистики.
2. Компьютерная технология создания словаря.
3. Построить семантический треугольник для слова *книга*.

В случае ведения учебного процесса с использованием электронного обучения и дистанционных образовательных технологий, содержание билета может отличаться от приведенного.

8. РАСПРЕДЕЛЕНИЕ БАЛЛОВ, КОТОРЫЕ ПОЛУЧАЮТ ОБУЧАЮЩИЕСЯ

Общая оценка знаний обучающихся по дисциплине проводится по 100-балльной шкале исходя из максимума, приведенного в таблице ниже.

Организационно-учебная работа в аудитории оценивается на основе таких критериев как посещаемость занятий, своевременное и качественное выполнение домашних заданий, активность во время проведения лекционных и практических занятий (участие в обсуждении текущего и пройденного материала, решение задач и т.п.).

Самостоятельная работа оценивается на основе предоставленных на проверку выполненных домашних, индивидуальных заданий с учетом своевременности их предоставления и соответствия требованиям к их выполнению.

Количество баллов за контрольную работу вычисляется как сумма баллов за все входящие в её состав задания. Каждое задание оценивается исходя из максимально возможного количества баллов с учетом правильности выполнения задания, полноты приводимых обоснований.

По результатам работы в семестре обучающийся, набравший не менее 60 баллов, имеет право получить оценку. Те, кто претендует на более высокий балл, проходят промежуточную аттестацию. Максимальное количество баллов на промежуточной аттестации – 100. Общее количество баллов за семестр вычисляется как максимальная из полученных за семестр и на промежуточной аттестации и выставляется согласно принятому порядку.

8.1. Семестр 5

Номера разделов	Виды работ	Максимальное количество баллов
1	Организационно-учебная работа в аудитории	10
	Самостоятельная работа	15
	Контрольная работа	35
Зачет		40
Общий итог за семестр		100

Семестр 6

Номера разделов	Виды работ	Максимальное количество баллов
2	Организационно-учебная работа в аудитории	10
	Самостоятельная работа	15
	Контрольная работа	35
Экзамен		40
Общий итог за семестр		100

Соответствие баллов оценке

Количество баллов из 100	ECTS	Оценка по пятибалльной шкале	
		Экзамен, дифференцированный зачет	Зачет
90-100	A	отлично	зачтено
80-89	B	хорошо	зачтено
75-79	C		зачтено
70-74	D	удовлетворительно	зачтено
60-69	E		зачтено
35-59	FX	неудовлетворительно	не зачтено
0-34	F		не зачтено

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОГО ПРОЦЕССА

Учебные занятия проводятся в 1-м учебном корпусе ДонГУ (г. Донецк, ул. Университетская, 24). Для проведения лекционных и практических занятий требуется аудитория, оборудованная меловой или маркерной доской, мультимедийный проектор и экран, ноутбук, комплект учебной мебели для студентов, рабочее место преподавателя, выход в Интернет – проводной или с использованием Wi-Fi.

Для самостоятельной работы используются текстовые и электронные ресурсы Научной библиотеки университета и других электронных библиотечных баз данных, учебно-методическое обеспечение, представленное на кафедре.

При изучении дисциплины применяются электронное обучение и дистанционные образовательные технологии.

10. РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

10.1. Основная литература

1. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. – М.: Изд-во НИУ ВШЭ, 2017.

2. Большакова и др. (2011). Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. – 272 с.

3. Математическая лингвистика и автоматическая обработка текстов : учеб. пособие / Т. В. Батура ; Новосиб. гос. ун-т. – Новосибирск : РИЦ НГУ, 2016 – 166 с.

4. Зубов А.В. Информационные технологии в лингвистике: учебное пособие для студентов вузов. – М.: Academia, 2004. – 205 с.

5. Хроленко А.Т., Денисов А.В. Грудева, Е.В. Корпусная лингвистика [Электронный ресурс] : учеб. пособие / Е. В. Грудева. – 2-е изд., стер. – М. : ФЛИНТА, 2012. – 165 с.

6. Обзор методов автоматической обработки текстов на естественном языке / С.Д.Белов, Д. П. Зрелова, П. В. Зрелов, В. В. Кореньков // Системный анализ в науке и образовании: сетевое научное издание. – 2020. – № 3. – С. 8–22.

10.1. Дополнительная литература

1. Автоматическая обработка текста. Графематика. <http://www.aot.ru/docs/graphan.html>.

2. Апресян Ю. Д. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы //Национальный корпус русского языка. – 2003. – Т. 2005. – С. 193-214.

3. Леонов В. П. О методах автоматического реферирования / В. П. Леонов // НТИ.Сер. 2 – 1975 – № 6 – С. 16–20.

4. Луканин А. В. Автоматическая обработка естественного языка/ А. В. Луканин; М-во образования и науки Российской Федерации, Южно-Уральский гос. ун-т. – Челябинск: Изд. центр ЮУрГУ, 2011 – 70 с.

5. Нейросетевые методы в обработке естественного языка / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2019. – 282 с.

11. ИНФОРМАЦИОННЫЕ РЕСУРСЫ

1. **Национальная электронная библиотека (НЭБ):** федеральная государственная информационная система / Министерство Культуры РФ; Российская государственная библиотека. – Москва, 2019- . – URL: <https://rusneb.ru/> (дата обращения: 31.03.2025). – Режим доступа: свободный, подписка. Необходима установка программного обеспечения. – Текст: электронный.

2. **eLIBRARY.RU:** научная электронная библиотека: сайт. – Москва, 2000- . – URL: <https://elibrary.ru> (дата обращения: 31.03.2025). – Режим доступа: для авторизов. пользователей. – Текст: электронный.

3. Научная электронная библиотека **«КиберЛенинка»:** сайт / Ассоциация «Открытая наука». – Москва, 2014- . – URL: <https://cyberleninka.ru/> (дата обращения: 31.03.2025). – Режим доступа: свободный. – Текст: электронный.

4. Электронно-библиотечная система **«Лань»:** [сайт]. – URL: <https://e.lanbook.com> (дата обращения: 31.03.2025). – Режим доступа: издания Сетевой электронной библиотеки, для авторизов. пользователей. – Текст: электронный.

5. **ЭБС Юрайт:** электронная библиотечная система: сайт. – Москва, 2013. – URL: <https://urait.ru/library/svobodnyy-dostup/> (дата обращения: 31.03.2025). – Режим доступа: издания свободного доступа, для авторизов. пользователей. – Текст: электронный.

6. **Электронно-библиотечная система ДонГУ:** сайт / ФГБОУ ВО «ДонГУ». – Донецк, 2016- . – URL: <http://library.donnu.ru/> (дата обращения: 31.03.2025). – Режим доступа: свободный. – Текст: электронный.

7. **Электронный каталог** Научной библиотеки ДонГУ: раздел сайта / НБ ДонГУ. – Текст: электронный // ЭБС ДонГУ: сайт. – URL: <http://library.donnu.ru/catalog/> (дата обращения: 31.03.2025). – Режим доступа: поиск свободный, электронные документы – для пользователей ДонГУ.

8. **Электронный архив ДонГУ:** раздел сайта / НБ ДонГУ. – Текст: электронный // ЭБС ДонГУ: сайт. – URL: <http://repo.donnu.ru/> (дата обращения: 31.03.2025). – Режим доступа: свободный.

12. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

1. Windows 7 PRO (корпоративная лицензия ДонГУ № 46484614)
2. Microsoft Office (корпоративная лицензия ДонГУ № 46472919)
3. Microsoft Visual Studio (лицензия программы Dream Spark для высших учебных заведений)
4. Антивирус Касперского, Adobe Acrobat Reader, xPDF (лицензии GPL, Apache, BSD для свободного программного обеспечения).