

УТВЕРЖДАЮ
 Декан факультета

_____ Шматко А.Д.

« ____ » _____ 20__

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Направление/специальность подготовки	45.05.01 Перевод и переводоведение
Специализация/профиль/программа подготовки	Лингвистика и современные цифровые технологии
Уровень высшего образования	Специалитет
Форма обучения	Очная
Факультет	Б Базовое инженерное образование
Выпускающая кафедра	Б5 Теоретическая и прикладная лингвистика
Кафедра-разработчик рабочей программы	Б5 Теоретическая и прикладная лингвистика

КУРС	СЕМЕСТР	ОБЩАЯ ТРУДОЁМКОСТЬ (ЗАЧЕТНЫХ ЕДИНИЦ)	ЧАСЫ (по наличию видов занятий)									ВИД ПРОМЕЖУТОЧНОГО КОНТРОЛЯ
			ОБЩАЯ ТРУДОЁМКОСТЬ	АУДИТОРНЫЕ ЗАНЯТИЯ				САМОСТОЯТЕЛЬНАЯ РАБОТА				
				ВСЕГО	ЛЕКЦИИ	ЛАБОРАТОРНЫЙ ПРАКТИКУМ	ПРАКТИЧЕСКИЕ ЗАНЯТИЯ	ВСЕГО	КУРСОВОЙ ПРОЕКТ	КУРСОВАЯ РАБОТА	ДРУГИЕ ВИДЫ САМОСТ. РАБОТЫ	
5	9	3	108	34	17	0	17	74	0	0	74	ЭКЗ.

ЛИСТ СОГЛАСОВАНИЯ

**РАБОЧАЯ ПРОГРАММА СОСТАВЛЕНА В СООТВЕТСТВИИ С ТРЕБОВАНИЯМИ ФЕДЕРАЛЬНОГО
ГОСУДАРСТВЕННОГО ОБРАЗОВАТЕЛЬНОГО СТАНДАРТА ВЫСШЕГО ОБРАЗОВАНИЯ (ФГОС ВО)**

45.05.01 Перевод и переводоведение

год набора группы: 2026

Программу составил:

Кафедра Б5 Теоретическая и прикладная лингвистика
Мамаев Иван Дмитриевич, к.ф.н., доцент

Программа рассмотрена
на заседании кафедры-разработчика
рабочей программы **Б5 Теоретическая и прикладная лингвистика**

Заведующий кафедрой Невзорова Г.Д., к.ф.н., доц.

Программа рассмотрена
на заседании выпускающей кафедры

Б5 Теоретическая и прикладная лингвистика

Заведующий кафедрой Невзорова Г.Д., к.ф.н., доц.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Разделы рабочей программы

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ
2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО
3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ
4. ФОРМЫ КОНТРОЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ
5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ
6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Приложения к рабочей программе дисциплины

- Приложение 1. Аннотация рабочей программы
- Приложение 2. Технологии и формы обучения
- Приложение 3. Фонды оценочных средств

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью освоения дисциплины является формирование следующих компетенций:

ПК-1 — Способен использовать различные цифровые средства, позволяющие достигать поставленных профессиональных целей

Формированию компетенций служит достижение следующих результатов образования:

ПК-1

знания:

знать архитектуру и методы работы с библиотеками обработки естественного языка;;

умения:

уметь настраивать и запускать пайплайны предобработки текста (очистка, нормализация, фильтрация стоп-слов) с использованием фреймворков машинного обучения;;

навыки:

обладать навыком разработки лингвистических правил для автоматического извлечения коллокаций, синтаксических шаблонов и терминологии из неструктурированных текстов на русском и английском языках..

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО

Дисциплина **КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА** является дисциплиной **обязательной части блока 1** программы подготовки по направлению *45.05.01 Перевод и переводоведение*.

Содержание дисциплины является логическим продолжением дисциплин: **МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА**.

Содержание дисциплины является основой для освоения дисциплин: **ЛОКАЛИЗАЦИЯ ЦИФРОВОГО КОНТЕНТА, ПОСТРЕДАКТИРОВАНИЕ МАШИННОГО ПЕРЕВОДА**.

Предварительные компетенции, сформированные у обучающегося до начала изучения дисциплины:

- УК-1 — Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины составляет 3 з.е., 108 ч.

3.1. Содержание (дидактика) дисциплины

КУРС	СЕМЕСТР	Наименование разделов и дидактических единиц	ВСЕГО	Аудиторные занятия в контактной форме			Самостоятельная работа студентов	Формируемая компетенция, %
				ВСЕГО	Лекции	Практические занятия		ПК-1
5	9	Раздел 1. Введение в компьютерную лингвистику. Теоретические, методологические и практические основы использования компьютерных технологий в современной лингвистике. История использования математических методов в зарубежной и отечественной лингвистике.	16	4	2	2	12	10
5	9	Раздел 2. Компьютерная морфология. Автоматический морфологический анализ. Методы построения морфологических моделей. Словарь А.А. Зализняка и его влияние на современные реализации компьютерных морфологических систем для русского языка. Первые этапы обработки текста: токенизация, лемматизация, частеречная разметка. Снятие морфологической неоднозначности.	18	6	3	3	12	10
5	9	Раздел 3. Компьютерный синтаксис. Способы представления синтаксической структуры предложения в системах обработки текста. Сочинительные и подчинительные синтаксические связи. Грамматика зависимостей и ее сложности. Грамматика непосредственно-составляющих и ее сложности. Формальные грамматики Н. Хомского. Способы снятия синтаксической неоднозначности. Синтаксис в модели «Смысл <=> Текст».	18	6	3	3	12	20
5	9	Раздел 4. Современные проблемы машинного перевода. История машинного перевода, отечественные и зарубежные системы. Обзор основных стратегий машинного перевода: перевод, основанный на правилах, статистический машинный перевод, нейронный машинный перевод.	18	6	3	3	12	20
5	9	Раздел 5. Другие направления компьютерной лингвистики. Визуализация лингвистических данных. Методы классификации (метод k-ближайших соседей, деревья решений и др.) и кластеризации (метод k-средних, DBSCAN и др.) текстовых данных. Создание диалоговых систем и чат-ботов. Автоматическое извлечение информации. Синтез и распознавание речи.	19	6	3	3	13	20
5	9	Раздел 6. Компьютерная семантика. Автоматическое назначение семантических ролей. Извлечение значения слова из окружающего контекста. Методы дистрибутивной семантики. Алгоритмы тематического моделирования как способ семантической компрессии текстовых данных и их виды: алгебраические, вероятностные, контекстуализированные.	19	6	3	3	13	20
Всего за 9 семестр			108	34	17	17	74	100
Всего по дисциплине			108	34	17	17	74	100

3.2. Аудиторный практикум

№ п/п	Номер и наименование раздела дисциплины	Тема практического занятия	Объем, ауд. часов
1	Раздел 1. Введение в компьютерную лингвистику.	Обсуждение основных этапов становление компьютерной лингвистики.	2
2	Раздел 2. Компьютерная морфология.	Практическая работа с современными морфологическими анализаторами (pymorphy2, mystem, AOT, rnnmorph и др.). Морфологическая разметка текстовых данных.	3
3	Раздел 3. Компьютерный синтаксис.	Сравнение современных синтаксических парсеров (Link Grammar, Stanza, AIIE и пр.). Синтаксический анализ неоднозначного русскоязычного предложения в NLTK. Разработка общей грамматики.	3
4	Раздел 4. Современные проблемы машинного перевода.	Сравнение современных систем машинного перевода, оценка их достоинств и недостатков. Выбор текста для домашнего задания.	3
5	Раздел 5. Другие направления компьютерной лингвистики.	Работа с no-code платформой Orange. Сбор индивидуальных текстовых данных. Обработка данных и их классификация и кластеризация.	3
6	Раздел 6.	Работа с алгоритмом word2vec. Построение дистрибутивной модели	3

	Компьютерная семантика.	собственной текстовой коллекции, выделение ассоциатов. Работа с алгоритмами тематического моделирования в облачной среде программирования Google Colab (ЯП Python): LDA, LSI, BERTopic и др. Выбор собственного текста для дистрибутивного и тематического анализа.	
Всего за 9 семестр			17

3.3. Самостоятельная работа студента (СРС)

№ п/п	Номер и наименование раздела дисциплины	Содержание учебного задания	Объем, часов
1	Раздел 1. Введение в компьютерную лингвистику.	Изучение теоретических материалов по теме раздела.	12
2	Раздел 2. Компьютерная морфология.	Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	12
3	Раздел 3. Компьютерный синтаксис.	Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	12
4	Раздел 4. Современные проблемы машинного перевода.	Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	12
5	Раздел 5. Другие направления компьютерной лингвистики.	Изучение теоретических материалов по теме раздела. Работа над домашним заданием. Подготовка к экзамену.	13
6	Раздел 6. Компьютерная семантика.	Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	13
Всего за 9 семестр			74

4. ФОРМЫ КОНТРОЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

СЕМЕСТР	НЕДЕЛИ СЕМЕСТРА																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
9		Диск., КР		ДЗ, КР		ДР			ДЗ, КР	ДР		ДЗ			КР	ДР	КР

Условные обозначения:

- ДР – диагностическая работа;
- Диск. – дискуссия;
- ДЗ – домашнее задание;
- КР – курсовая работа.

Текущий контроль успеваемости студентов проводится в дискретные временные интервалы в следующих формах:

- диагностическая работа;
- дискуссия;
- домашнее задание;
- курсовая работа.

Промежуточная аттестация проводится в формах:

- экзамен.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература по дисциплине:

1. А. Н. Гуцин. . Технология обработки текста и звучащей речи. СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2018, 66 экз.
2. С. А. Гашков. . Формальные модели в лингвистике. СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2017, 58 экз.
3. С. А. Гашков. . Лингвистическая семантика. СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2017, 48 экз.
4. Я. Г. Тестелец. . Введение в общий синтаксис. М.: Изд-во РГГУ, 2001, эл. рес.

5.2. Дополнительная литература по дисциплине:

1. А. Н. Баранов. . Введение в прикладную лингвистику. М.: Эдиториал УРСС, 2003, 2 экз.

5.3. Периодические издания:

1. Моделирование и анализ информационных систем.

5.4. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины, электронные библиотечные системы:

1. <http://library.voenmeh.ru/jirbis2> — Сайт фундаментальной библиотеки БГТУ «Военмех» им. Д.Ф. Устинова — Фундаментальная библиотека БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова;
2. <https://e.lanbook.com/> — ЭБС Лань;
3. <https://urait.ru/> — Образовательная платформа «Юрайт». Для вузов и ссузов.

Современные профессиональные базы данных:

1. <https://rusneb.ru> – Национальная электронная библиотека (НЭБ);
2. <https://cyberleninka.ru/> - Научная электронная библиотека «Киберленинка»;
<http://www.rfbr.ru/rffi/ru/library> - Полнотекстовая электронная библиотека Российского фонда фундаментальных исследований.

Информационные справочные системы:

1. Техэксперт – Информационный портал технического регулирования: Нормы, правила, стандарты РФ;
2. http://library.voenmeh.ru/jirbis2/index.php?option=com_irbis&view=irbis&Itemid=457 - БД ГОСТов собственной генерации БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова;
3. <http://www.consultant.ru/>- КонсультантПлюс- информационный портал правовой информации.

5.5. Программное обеспечение:

1. Python 3.4.

5.6. Информационные технологии:

взаимодействие с обучающимися посредством ЭИОС Moodle БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Лекционные занятия:

специализированные требования по оборудованию отсутствуют; аудитория с посадочными местами по количеству студентов; доска.

6.2. Практические занятия:

1. Проектор;
2. Интерактивная доска;
3. Аудитория с числом посадочных мест не меньше количества обучающихся;
4. Python 3.4.

6.3. Прочее:

1. рабочее место преподавателя, оснащенное компьютером с доступом в Интернет;
2. рабочие места студентов, оснащенные компьютерами с доступом в Интернет, предназначенные для работы в электронной образовательной среде.

Аннотация рабочей программы

Дисциплина **КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА** является дисциплиной **обязательной части блока 1** программы подготовки по направлению *45.05.01 Перевод и переводоведение*. Дисциплина реализуется на факультете *Б* Базовое инженерное образование БГТУ "ВОЕНМЕХ" им. Д.Ф. Устинова кафедрой Б5 Теоретическая и прикладная лингвистика.

Дисциплина нацелена на формирование *компетенций*:

ПК-1 Способен использовать различные цифровые средства, позволяющие достигать поставленных профессиональных целей.

Содержание дисциплины охватывает круг вопросов, связанных с базами данных в сфере интеллектуальных технологий и прикладной лингвистики, а также с их практическим применением.

Программой дисциплины предусмотрены следующие **виды контроля**:

Текущий контроль успеваемости студентов проводится в дискретные временные интервалы в следующих формах:

- диагностическая работа;
- дискуссия;
- домашнее задание;
- курсовая работа.

Промежуточная аттестация проводится в формах:

- экзамен.

Общая трудоемкость освоения дисциплины составляет **3 з.е., 108 ч.** Программой дисциплины предусмотрены лекционные занятия (**17 ч.**), практические занятия (**17 ч.**), самостоятельная работа студента (**74 ч.**).

ТЕХНОЛОГИИ И ФОРМЫ ОБУЧЕНИЯ

Рекомендации по освоению дисциплины для студента

Трудоемкость освоения дисциплины составляет 108 ч., из них 34 ч. аудиторных занятий, и 74 ч., отведенных на самостоятельную работу студента.

Рекомендации по распределению учебного времени по видам самостоятельной работы и разделам дисциплины приведены в таблице.

Контроль освоения дисциплины производится в соответствии с Положением о текущем, рубежном контроле успеваемости и промежуточной аттестации обучающихся.

Формы контроля и критерии оценивания приведены в приложении 3 к Рабочей программе.

Наименование работы	Рекомендуемая литература	Трудоемкость, час.
Раздел 1. Введение в компьютерную лингвистику.		
Изучение теоретических материалов по теме раздела.	С. А. Гашков. . Формальные модели в лингвистике: СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2017 (1-7) А. Н. Баранов. . Введение в прикладную лингвистику: М.: Эдиториал УРСС, 2003 (1-6)	12
Итого по разделу 1		12
Раздел 2. Компьютерная морфология.		
Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	С. А. Гашков. . Формальные модели в лингвистике: СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2017 (1-7)	12
Итого по разделу 2		12
Раздел 3. Компьютерный синтаксис.		
Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	Я. Г. Тестелец. . Введение в общий синтаксис: М.: Изд-во РГГУ, 2001 (10-17)	12
Итого по разделу 3		12
Раздел 4. Современные проблемы машинного перевода.		
Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	А. Н. Гуцин. . Технология обработки текста и звучащей речи: СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2018 (17)	12
Итого по разделу 4		12
Раздел 5. Другие направления компьютерной лингвистики.		
Изучение теоретических материалов по теме раздела. Работа над домашним заданием. Подготовка к экзамену.	А. Н. Гуцин. . Технология обработки текста и звучащей речи: СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2018 (11-17)	13
Итого по разделу 5		13
Раздел 6. Компьютерная семантика.		
Изучение теоретических материалов по теме раздела. Работа над домашним заданием.	С. А. Гашков. . Лингвистическая семантика: СПб.БГТУ "ВОЕНМЕХ" им. Д. Ф. Устинова, 2017 (11)	13
Итого по разделу 6		13

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Фонд оценочных средств, позволяющие оценить результаты обучения по данной дисциплине, включают в себя:

- диагностическая работа
- дискуссия;
- домашнее задание;
- курсовая работа;
- экзамен.

Критерии оценивания

Диагностическая работа

Диагностическая работа проводится в форме теста в ЭИОС Moodle:

- при правильном ответе менее чем на 60% вопросов - не аттестация;
- при правильном ответе на 60% вопросов и более - аттестация.

Дискуссия

Дискуссия предполагает активное обсуждение современных теоретических направлений по компьютерной лингвистике, а также ответы на вопросы преподавателя. Перечень дискуссионных тем представлен ниже.

1. Общие направления компьютерной лингвистики.
2. История применения математических методов для лингвистического анализа.
3. Морфосинтаксические анализаторы.
4. Современный автоматический семантический анализ.
5. Лингвистические проблемы систем машинного перевода.

Домашнее задание

Все домашние задания выполняются по индивидуальным наборам текстовых данных. Индивидуальные наборы текстовых данных для обработки заранее оговариваются с преподавателем, он же фиксирует их в своём журнале. При выполнении домашнего задания студент должен продемонстрировать знание теоретического материала, относящегося к теме данной работы, обосновать целесообразность выбранных решений. Отчет по каждому домашнему заданию представляется в pdf-формате с листингом программы или ссылкой на неё, если это необходимо. Отчет не может быть принят и подлежит доработке в случае:

- отсутствия необходимых разделов,
- несодержательной интерпретации полученных данных,
- отсутствия необходимого графического материала,
- некорректного обоснования выбранных решений.

Курсовая работа

Перечень тем курсовых работ формируется на основе обсуждаемых в начале семестра тем. Оценка «отлично» выставляется за глубокое исследование проблемы, самостоятельные выводы и безупречное оформление работы. Студент должен уверенно доложить материал, свободно отвечать на вопросы и аргументированно защищать свою позицию. Для оценки «хорошо» допускается наличие незначительных неточностей в теории или оформлении, не влияющих на общую суть. Защита проходит успешно, если студент демонстрирует твердые знания, но иногда затрудняется с развернутыми ответами на сложные вопросы. Оценка «удовлетворительно» ставится при поверхностном анализе темы и наличии существенных замечаний по структуре или содержанию. На защите студент излагает материал по плану, но отвечает на вопросы комиссии с помощью наводящих подсказок. Работа считается неудовлетворительной, если тема раскрыта неполно, использованы устаревшие данные или выявлен плагиат. Отсутствие навыков презентации материала и неспособность ответить на базовые вопросы также ведут к отрицательному результату. Критерием качества служит соответствие работы методическим рекомендациям и актуальность предложенных решений. Итоговая оценка формируется на основе комплексного анализа письменного текста и устного выступления студента.

Экзамен

Оценка за экзамен выставляется на основании баллов, которые обучающийся набрал в течение семестра по балльно-рейтинговой системе. Процедура выставления баллов регламентируется на первых занятиях семестра. При несогласии с выставленной оценкой студент имеет право сдать экзамен в традиционной устной форме, ответив на два вопроса. Ниже приведен приблизительный список вопросов по теме дисциплины.

1. История и основные этапы развития компьютерной лингвистики.
2. Задачи морфологической обработки.
3. Морфологическая неоднозначность и методы ее разрешения в задачах компьютерной лингвистики.
4. Методы синтаксического парсинга: алгоритмы Кока-Касами-Янгера, Эрли и др. системы.
5. Синтаксическая неоднозначность и стратегии ее разрешения (по А.В. Гладкому).
6. Стандарт Universal Dependencies.
7. История машинного перевода.
8. Архитектура систем машинного перевода.
9. Лингвистические проблемы машинного перевода. Постредактирование текстов.
10. Дистрибутивная семантика. Широкая и узкая трактовки дистрибутивной гипотезы.
11. Векторные модели языка: word2vec, BERT, ELMo.
12. Тематическое моделирование текстов на естественном языке.
13. Оценка качества лингвистических систем.
14. Автоматическая кластеризация языковых данных.
15. Автоматическая классификация языковых данных.
16. Большие языковые модели. Стратегии промт-инжиниринга.
17. Распознавание именованных сущностей.
18. Тональный анализ текстов (сентимент-анализ).
19. Выделение ключевых слов.
20. Алгоритмы автоматической суммаризации текстов.
21. Лингвистические онтологии и инженерия знаний.

Паспорт фонда оценочных средств

КУРС	СЕМЕСТР	Наименование разделов и дидактических единиц	ВСЕГО	Аудиторные занятия в контактной форме			Самостоятельная работа студентов	Формируемая компетенция, %	НАИМЕНОВАНИЕ ОЦЕНОЧНОГО СРЕДСТВА
				ВСЕГО	Лекции	Практические занятия		ПК-1	
5	9	Раздел 1. Введение в компьютерную лингвистику.	16	4	2	2	12	10	Дискуссия
5	9	Раздел 2. Компьютерная морфология.	18	6	3	3	12	10	Домашнее задание, Курсовая работа
5	9	Раздел 3. Компьютерный синтаксис.	18	6	3	3	12	20	Домашнее задание
5	9	Раздел 4. Современные проблемы машинного перевода.	18	6	3	3	12	20	Домашнее задание
5	9	Раздел 5. Другие направления компьютерной лингвистики.	19	6	3	3	13	20	Домашнее задание
5	9	Раздел 6. Компьютерная семантика.	19	6	3	3	13	20	Домашнее задание, Курсовая работа
Всего за 9 семестр			108	34	17	17	74	100	
Всего по дисциплине			108	34	17	17	74	100	

Оценочные материалы по дисциплине КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

ПК-1 - Способен использовать различные цифровые средства, позволяющие достигать поставленных профессиональных целей

№ 1 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов

Что из нижеперечисленного является способами нейросетевой векторизации некоторой модели текста?

1. Word2Vec.
2. TF-IDF.
3. BERT.
4. GloVe.
5. FastText.

№ 2 Прочитайте текст и запишите развернутый обоснованный ответ

Приведите пример синтаксической неоднозначности на английском языке и способ ее разрешения.

№ 3 Прочитайте текст и запишите развернутый обоснованный ответ

Объясните, почему токенизация для китайского или японского языка сложнее, чем для английского.

№ 4 Прочитайте текст и установите соответствие

Установите соответствие между типами языковых моделей и их характеристиками.

МОДЕЛИ

ХАРАКТЕРИСТИКИ

- | | |
|------------------|----------------------------------------------------------|
| 1) N-gram models | A) Учитывают дальние зависимости, требуют много ресурсов |
| 2) RNN/LSTM | B) Основаны на частоте последовательностей слов |
| 3) Transformer | C) Обработывают последовательности рекуррентно |
| | D) Игнорируют порядок слов, только частоты |

№ 5 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа

Как называются онтологии, основанные на принципах меронимии и холонимии?

1. Партономии.
2. Словарь.
3. Таксономии.
4. Нет правильного ответа.

№ 6 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов

Какие архитектуры используются в современном машинном переводе?

1. Encoder-Decoder.
2. Random Forest.
3. Transformer.
4. K-means clustering.
5. Attention mechanism.

№ 7 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа

Какая метрика наиболее часто используется для оценки качества машинного перевода?

1. Accuracy.
2. BLEU.
3. F1-score.
4. Silhouette coefficient.

№ 8 Прочитайте текст, выберите правильные ответы и запишите аргументы, обосновывающие выбор ответов

Какие из перечисленных задач относятся к морфологическому анализу?

1. Определение части речи слова.
2. Построение дерева зависимостей.
3. Лемматизация.
4. Распознавание именованных сущностей (NER).
5. Определение падежа, числа, рода.

№ 9 Прочитайте текст и установите соответствие

Установите соответствие между задачами NLP и методами их решения.

ЗАДАЧИ

МЕТОДЫ

- | | |
|-------------------------------------------|------------------------------------|
| 1)
Синтаксический парсинг | A) CRF (Conditional Random Fields) |
| 2)
Распознавание именованных сущностей | B) Алгоритм СКУ |
| 3)
Классификация языковых данных | C) Beam Search |
| | D) Logistic Regression |

№ 10 Прочитайте текст и установите последовательность

Установите правильную последовательность этапов предобработки русскоязычного текста в NLP-пайплайне.

1. Лемматизация.
2. Токенизация.
3. Удаление стоп-слов, представленных в канонической форме.
4. Предварительное приведение слов текста в нижний регистр.

№ 11 Прочитайте текст и установите последовательность

Установите правильную последовательность развития подходов к машинному переводу (от ранних к современным).

1. Нейросетевой машинный перевод (NMT).
2. Машинный перевод на основе правил (RBMT).
3. Статистический машинный перевод (SMT).
4. Примеро-ориентированный перевод (EBMT).

№ 12 Прочитайте текст, выберите правильный ответ и запишите аргументы, обосновывающие выбор ответа

Какой метод используется для приведения слова к его начальной форме с учетом части речи?

1. Стемминг.
2. Лемматизация.
3. Токенизация.
4. N-граммы.