

Министерство науки и высшего образования Российской Федерации  
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

САЕ «Институт человека цифровой эпохи»

УТВЕРЖДАЮ:  
Руководитель ОПОП:



З.И. Резанова

« 31 » августа 20 22 г.

Рабочая программа дисциплины

**Корпусная лингвистика: создание и использование корпусов**

по направлению подготовки

**45.04.03 Фундаментальная и прикладная лингвистика**

Направленность (профиль) подготовки :  
**Компьютерная и когнитивная лингвистика**

Форма обучения  
**Очная**


Квалификация  
**Магистр**

Год приема  
**2022**

Код дисциплины в учебном плане: Б1.В.ДВ.2.5

СОГЛАСОВАНО:

Руководитель ОПОП

 З.И. Резанова

Председатель УМК

 Ю.А. Тихомирова

- **1. Цель и планируемые результаты освоения дисциплины**

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-1 – способен решать профессиональные задачи, применяя основные понятия, категории и положения лингвистических теорий и актуальные концепции в области лингвистики.

ОПК-2 – способен анализировать, сопоставлять и критически оценивать различные лингвистические направления, теории и гипотезы при решении задач профессиональной деятельности.

ОПК-3 – способен выбирать оптимальные подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий.

ПК-3 – способен разрабатывать системы автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистические компоненты электронных ресурсов и интеллектуальных электронных систем (лингвистические корпуса, словари, онтологии, базы данных).

ПК-4 – способен разрабатывать проекты прикладной направленности в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-1.2 – решает профессиональные задачи применяя основные понятия, категории и положения лингвистических теорий.

ИОПК-2.2 – формулирует гипотезы в рамках предложенного лингвистического направления и теории при решении задач профессиональной деятельности.

ИОПК-2.3 – совершает выбор лингвистического направления, теории на основе их самостоятельного поиска и анализа, сопоставления, критической оценки при решении задач профессиональной деятельности.

ИОПК-3.2 – критически сопоставляет и оценивает существующие подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий.

ИПК-3.2 – разрабатывает лингвистические компоненты электронных ресурсов (лингвистические корпуса, словари).

ИПК-4.1 – формулирует цель проекта прикладной направленности в области когнитивной и компьютерной лингвистики, обосновывает необходимость применения современных технических средств и информационных технологий, в том числе в области искусственного интеллекта.

ИПК-4.3 – обеспечивает выполнение проекта в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта, в соответствии с установленными целями, сроками и затратами.

- **2. Задачи освоения дисциплины**

- Формирование умений и навыков разработки лингвистических компонентов лингвистических корпусов.

- Формирование умений в сфере разработки и совершенствования системы автоматизации и информационной поддержки корпусов текстов.

- **3. Место дисциплины в структуре образовательной программы**

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор.

- **4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине**  
Третий семестр, зачет

- **5. Входные требования для освоения дисциплины**

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: Введение в корпусную лингвистику. Лингвистика в контексте современного гуманитарного и естественно-научного знания. Введение в анализ естественного языка (NLP).

- **6. Язык реализации**

Русский

- **7. Объем дисциплины**

Общая трудоемкость дисциплины составляет 2 з.е., 72 часа, из которых:

лекции: 6 ч;

практические занятия: 22 ч;

Объем самостоятельной работы студента определен учебным планом.

- **8. Содержание дисциплины, структурированное по темам**

Тема 1. Ключевые аспекты создания корпусов разных типов. Схема аннотирования корпусов разных типов.

Тема 2. Зачем нужны корпусные данные. Частотность. Ключевые слова.

Использование корпусного менеджера для представления корпуса и работы с корпусными данными.

Тема 3. Корпус русской устной речи тюркско-русских билингвов. Структура корпуса и принципы разметки. Типы решаемых задач. Инструменты создания и аннотирования корпуса устной речи.

Тема 4. Корпусная лингвистика и NLP. Автоматическая разметка.

Тема 5. Создание и аннотирование специализированного корпуса.

Тема 6. Использование корпуса в исследовании дискурса.

- **9. Текущий контроль по дисциплине**

Текущий контроль по дисциплине проводится путем контроля посещаемости, выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

- **10. Порядок проведения и критерии оценивания промежуточной аттестации**

Зачет с оценкой в третьем семестре принимается в форме проекта.

Для зачета необходимо выполнить следующее задание:

Представление фрагмента проекта лингвистического корпуса текстов, отвечающего перечисленным требованиям.

Проект разработан как целостная концепция; обоснованы и соотнесены: цель корпуса, принципы отбора материала, принципы метаразметки и лингвистического аннотирования. Представлен репрезентативный фрагмент материалов корпуса, вариант метаразметки и разметки произведен лингвистически грамотно, обоснование соответствует современным лингвистическим концепциям, представлен фрагмент автоматического аннотирования.

Проект проверяет компетенции ИОПК-1.2, ИОПК-2.2, ИОПК-2.3, ИОПК-3.2, ИПК-3.2, ИПК-4.1, ИПК-4.3.

Результаты зачета с оценкой определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценка «отлично» ставится при своевременной и успешной сдаче проекта, своевременном выполнении домашних заданий и посещаемости не менее 80% занятий.

Оценка «хорошо» ставится при успешной сдаче проекта и выполнении домашних заданий, посещаемости не менее 60% занятий.

Оценка «удовлетворительно» ставится при сдаче проекта и выполнении домашних заданий при посещаемости менее 60% занятий.

Оценка «неудовлетворительно» ставится при несдаче проекта или невыполнении домашних заданий.

- **11. Учебно-методическое обеспечение**

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - <https://moodle.tsu.ru/course/view.php?id=3590>

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

в) Методические указания по организации самостоятельной работы студентов.

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

В приведенной ниже таблице представлены методические указания по организации самостоятельной работы студентов по темам.

Тема	Самостоятельная работа
Тема 1. Ключевые аспекты создания корпусов разных типов. Схема аннотирования корпусов разных типов.	<p>Студенты знакомятся с интерфейсом среды разработки, учатся писать команды и запускать их, назначать переменные и подавать аргументы в функции.</p> <p>Самостоятельная работа: Чтение статьи М. McCarthy and A.O’Keeffe, Historical perspective. What are corpora and how have they evolved?</p>
Тема 2. Зачем нужны корпусные данные. Частотность. Ключевые слова. Использование корпусного менеджера для представления корпуса и работы с корпусными данными.	<p>Каждый студент работает с данными в корпусном менеджере.</p> <p>Самостоятельная работа: ознакомиться с документацией к корпусному менеджеру на выбор: LancsBox или AntConc. (<a href="http://corpora.lancs.ac.uk/lancsbox/help.php">http://corpora.lancs.ac.uk/lancsbox/help.php</a>, <a href="https://www.laurenceanthony.net/software/antconc/">https://www.laurenceanthony.net/software/antconc/</a>)</p> <p>Brezina, V. Statistics in corpus linguistics : a practical guide. Chapters 2-3.</p>
Тема 3. Корпус русской устной речи тюркско-русских билингов. Структура корпуса и принципы разметки. Типы решаемых задач. Инструменты создания и аннотирования корпуса устной речи.	<p>Студенты учатся работать с программой ELAN. Изучение принципов транскрипции и аннотирования данных устной речи. Лингвистическая разметка. Метаразметка.</p> <p>Самостоятельная работа: ознакомление с инструкциями по сбору данных и разметке корпуса. Изучение документации к программе ELAN. (<a href="https://archive.mpi.nl/tla/elan/documentation">https://archive.mpi.nl/tla/elan/documentation</a>)</p>
Тема 4. Корпусная лингвистика и NLP. Автоматическая разметка.	<p>Студенты знакомятся с применением методов и инструментов корпусной лингвистики в NLP. Сравнение инструментов автоматической морфологической разметки.</p> <p>Самостоятельная работа: Ознакомление с документацией морфологических анализаторов (<a href="https://pymorphy2.readthedocs.io/en/stable/">https://pymorphy2.readthedocs.io/en/stable/</a>, <a href="https://yandex.ru/dev/mystem/">https://yandex.ru/dev/mystem/</a>,</p> <p>Hammond M. Python for Linguists / M. Hammond. – Cambridge University Press, 2020. – p. 41</p>
Тема 5. Создание и аннотирование специализированного корпуса.	<p>Студенты создают небольшой специализированный корпус на выбор: по теме магистерской диссертации или фрагмент корпуса устной речи.</p> <p>Захаров В.П., Азарова И.В., Митрофанова О.А., Попов А.М., Хохлова М.В.: Моделирование в корпусной лингвистике: специализированные корпуса русского языка. СПб.: Изд-во С.-Петербур. ун-та, 2019.</p>
Тема 6. Использование корпуса в исследовании дискурса.	<p>Знакомство с примерами использования корпуса исследования дискурса.</p> <p>Love R., Dembry C., Hardie A., Brezina V., McEneaney T. Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, Volume 22, Issue 3, Nov 2017, p. 319 — 344.</p> <p>Atkins S., Harvey K. The Routledge Handbook of Corpus Linguistics, Chapter How to use corpus linguistics in the study of health communication.</p> <p>Захаров В. П., Богданова С. Ю. Корпусная</p>

	лингвистика: учебник. 3-е изд., перераб. Глава 12. Исследования дискурса, основанные на корпусах.
--	---

● **12. Перечень учебной литературы и ресурсов сети Интернет**

а) основная литература:

– Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. / В.П. Захаров [и др.]. – СПб.: Изд-во С.-Петербур. ун-та, 2020. – 234 с.

– Моделирование в корпусной лингвистике: специализированные корпуса русского языка / В.П.Захаров, И.В.Азарова, О.А.Митрофанова, А.М.Попов, М.В.Хохлова; отв. ред. В.П.Захаров . – СПб.: Изд-во С.-Петербур. Ун-та, 2019. — 208 с.

б) дополнительная литература:

– Rezanova Z. I., Temnikova I. G. Artemenko E. D. Stepanenko A. A. Dat sy uk V. V. Dybo A. V. THE BIMODAL CORPUS OF RUSSIAN-TURKIC BILINGUALS' SPEECH (RuTuBiC) // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог» (2019). Выпуск 18. Дополнительный том. С. 200-210. [http://www.dialog-21.ru/media/4870/\\_-dialog2019scopusvolplus.pdf](http://www.dialog-21.ru/media/4870/_-dialog2019scopusvolplus.pdf)

– Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003—2005. М.:Индрик, 2005, 193—214.

– Гришина Е. А., Савчук С. О. Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 129—149.

– Зобнин А. И., Сахарова А. В. Универсальная система разметки текста ObjectATE // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 283—297.

– Рахилина Е. В. О новых инструментах описания русской грамматики: корпус ошибок // Русский язык за рубежом. 2016. № 3. С. 20-25

– Резникова Т. И., Копотев М. В. Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов)// Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 31—61.

– Сичинава Д. В. Обработка текстов с грамматической разметкой: инструкция разметчика // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 136—154.

в) ресурсы сети Интернет:

– Национальный корпус русского языка// <http://www.ruscorpora.ru>

– The Russian Learner Corpus (RLC) URL: <http://web-corpora.net/RLC>

- EAGLES. Preliminary recommendations on Corpus Typology. EAG—TCWG—СТУП/P. Version of May, 1996 // URL: <http://www.ilc.cnr.it/EAGLES/corpus/typ/corpus.html>
- Корпус контактно-обусловленной русской речи билингвов-носителей малых языков Севера Сибири и Дальнего Востока. URL: <http://web-corpora.net/ruscontact/>
- Словари, созданные на основе Национального корпуса русского языка. URL: <http://dict.ruslang.ru/>

### ● 13. Перечень информационных технологий

- а) лицензионное и свободно распространяемое программное обеспечение:
  - Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);
  - публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).
  - публично доступный редактор лингвистического аннотирования ELAN URL: <https://archive.mpi.nl/tla/elan>
  - публично доступная программа морфологического анализа MyStem. URL: <https://yandex.ru/dev/mystem/>
  - публично доступный корпусный менеджер AntConc URL: <https://www.laurenceanthony.net/software/antconc/>
  - публично доступный корпусный менеджер Lancs Box URL: <http://corpora.lancs.ac.uk/lancsbox/index.php>

- б) информационные справочные системы:
  - Электронный каталог Научной библиотеки ТГУ – <http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system>
  - Электронная библиотека (репозиторий) ТГУ – <http://vital.lib.tsu.ru/vital/access/manager/Index>
  - ЭБС Лань – <http://e.lanbook.com/>
  - ЭБС Консультант студента – <http://www.studentlibrary.ru/>
  - Образовательная платформа Юрайт – <https://urait.ru/>
  - ЭБС ZNANIUM.com – <https://znanium.com/>
  - ЭБС IPRbooks – <http://www.iprbookshop.ru/>

- в) профессиональные базы данных (*при наличии*):
  - Университетская информационная система РОССИЯ – <https://uisrussia.msu.ru/>
  - Единая межведомственная информационно-статистическая система (ЕМИСС) – <https://www.fedstat.ru/>
  - Общее языкознание [Электронный ресурс] // Philology.ru: Русский филологический портал. – URL: [http://www.philology.ru/linguistics1.htm\\_](http://www.philology.ru/linguistics1.htm_)
  - Языкознание [Электронный ресурс] // Библиотека Гумер. – URL: [http://www.gumer.info/bibliotek\\_Buks/Linguist/Index\\_Lin](http://www.gumer.info/bibliotek_Buks/Linguist/Index_Lin)
  - Serious-science. Linguistics (Серьезная наука. Лингвистика) [Электронный ресурс]. – URL: <http://serious-science.org/themes/linguistics>.
  - Serious-science. Linguistics. (Серьезная наука. Лингвистика)[Электронный ресурс] URL: <http://serious-science.org/themes/linguistics>
  - Scopus: database[Электронный ресурс]. URL: <https://www.scopus.com/> (01.09.2016).
  - Web of Science: database [Электронный ресурс]. URL: <http://login.webofknowledge.com/> (01.09.2016).

- Elibrary.ru: научная электронная библиотека [Электронный ресурс]. URL: [http://elibrary.ru/project\\_risc.asp](http://elibrary.ru/project_risc.asp) (01.09.2016).
- Научно-образовательный портал «Лингвистика в России: ресурсы для исследователей» [Электронный ресурс]. URL: <http://uisrussia.msu.ru/linguist/index.jsp>
- Научно-образовательный портал «Лингвистика в России: ресурсы для исследователей» [Classes.ru. Иностранные языки для всех. Словари онлайн. [Электронный ресурс]. URL: <http://www.classes.ru>

- **14. Материально-техническое обеспечение**

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

- **15. Информация о разработчиках**

Резанова Зоя Ивановна, д-р филол. наук, профессор, НИ ТГУ, зав. каф. общей, компьютерной и когнитивной лингвистики

Погодаева Елена Николаевна, НИ ТГУ, ассистент каф. общей, компьютерной и когнитивной лингвистики