

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

Филологический факультет

УТВЕРЖДЕНО:
Декан
И. В.Тубалова

Рабочая программа дисциплины

Технологии корпусной лингвистики

по направлению подготовки

45.03.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки:
Фундаментальная и прикладная лингвистика

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
А.В. Васильева

Председатель УМК
Ю.А. Тихомирова

Томск – 2024

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

- ПК-4 Способен разрабатывать программный код при решении задач автоматической обработки текстов;
- 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

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

- ИОПК-2.3 Применяет математические методы в процедурах анализа и синтеза лингвистических структур
- ИПК-4.1 Применяет способы формализации и алгоритмизации поставленных задач в сфере автоматической обработки текстов
- ИПК-4.2 Использует базовые методы и приемы различных типов устной и письменной коммуникации на русском литературном языке в профессиональной деятельности

2. Задачи освоения дисциплины

- Освоить математический аппарат анализа языковых явлений на базе технологий корпусной лингвистики.
- Разрабатывать корпус, знать и понимать основные принципы работы корпусной лингвистики.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к обязательной части образовательной программы.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Шестой семестр, экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: информатика и основы программирования, математическая статистика, информационные технологии и основы информационной культуры в лингвистике.

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 6 з.е., 144 часов, из которых:

-лекции: 20 ч.

-практические занятия: 30 ч.

в том числе практическая подготовка: 30 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Введение в корпусную лингвистику

Введение в корпусную лингвистику

Типология корпусов

Разметка корпуса: лингвистическая и экстралингвистическая разметки
Основные свойства корпуса

Тема 2. Практика по работе в наиболее известных корпусах
Теоретические и практические аспекты использования НКРЯ
Изучение web-интерфейса и функций корпусов: НКРЯ, КРУТ, RLC и пр.

Поиск и анализ n-грамм

Конкордансный поиск ЛЕ в корпусе

Семантический поиск

Тема 3. Квантитативные методы исследования

Принципы выбора и обоснование использования квантитативных методов

Абсолютная частота. Относительная частота

Коэффициент Жуйана.

Коэффициент логарифмического правдоподобия LL-score

Создание базы данных

Критерий отношения правдоподобия (loglikelihood)

Метрики для определения ключевых слов

Самостоятельная работа по теории

Тема 4. Разметка корпуса

Основные принципы и типы разметки корпуса: лингвистическая и экстралингвистическая разметки

Морфологическая разметка корпуса: mystem, AOT, TreeTagger, Marmot

Синтаксическая разметка корпуса: UDPIPE

Аннотированная разметка корпуса с помощью ELAN

Тема 5. Разработка корпуса

Принципы разработки корпуса

Парсинг текстов с помощью BootCat

Сборка корпуса в ПО AntConc

Анализ разработанного корпуса

9. Текущий контроль по дисциплине

Текущий контроль образовательной программы (темы, раздела, модуля) требованиям образовательных стандартов по направлениям подготовки/специальностям. Текущий контроль успеваемости обучающихся направлен на определение соответствия результатов обучения после освоения элемента по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, тестов по лекционному материалу, разработки кода, выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр. Примерные задания текущего контроля:

1. Найдите корпус языка, который вы изучаете (Например, если вы учите английский - <https://www.english-corpora.org/bnc/>)

2. Сравните частоту использования слова «информация» в корпусах НКРЯ и BNC (подкорпус литература и СМИ)

3. Сравните, используя формулу IPM, два слова в подкорпусах найденного вами корпуса и НКРЯ. Результат анализа оформить с помощью формул в Excel. Добавьте в отдельный лист сравнительную гистограмму.

4. Сравните IPM (instances per million words) двух слов в разных корпусах или подкорпусах. Варианты сравнения:

а) одно и то же слово в двух разных подкорпусах НКРЯ (например, в основном и в газетном). Интересным будет сравнение синонимов или антонимов в одном подкорпусе и разных подкорпусах.

б) одно и то же слово в двух разных подкорпусах (например, www.english-corpora.org/bnc/ -> comparisons between genres)

с) слово на русском в НКРЯ и его перевод английский в (<https://www.english-corpora.org/coca/> или <https://www.english-corpora.org/bnc/>)

d) два разных слова в любом корпусе изучаемого иностранного языка.

Оформите результаты в таблице Excel, вставьте графики на основе таблицы. Образцы оформления и вычисления см. в файле IPM_TF-IDF.xlsx

Можно выбрать один из двух вариантов оформления на одном из листов, но можно оформить по-своему. Наличие вычислительных формул в ваших таблицах - обязательно.

5. Сравните статистическую меру TF-IDF у двух текстов из НКРЯ, статистически оцените важность слова (лексемы) в контексте документа.

Для повторения метода TF-IDF обратите внимание на презентацию TF-IDF.ppt с главной страницы курса.

Выберите поиск одного слова в любом подкорпусе НКРЯ (или в двух разных подкорпусах). Слово вводите в поисковой форме "Лексико-грамматический поиск", а не в поиске точных форм. Никаких дополнительных грамматических ограничений устанавливать не нужно. Сравните важность этого слова в двух разных текстах. Количество словоформ в тексте можно увидеть в перечне параметров текста, если кликнуть по названию текста.

Образцы оформления и вычисления см. в файле IPM_TF-IDF.xlsx

Обязательно приложите и вставьте под соответствующими таблицами скриншоты с метатекстовой информацией (см. в образце) для подтверждения достоверности числовых данных из текстов.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзамен в шестом семестре проводится в письменной форме по билетам. Экзаменационный билет состоит из трех частей. Продолжительность экзамена 1,5 часа.

Первая часть представляет собой ответ на теоретический вопрос в устной форме, проверяющую компетенции ИОПК-2.3; ИПК-4.1; ИПК-4.2. Ответы на вопросы первой части даются путем выбора из списка предложенных.

Примерный перечень теоретических вопросов:

1. Дайте определение корпусной лингвистики?
2. Что означает формула IPM и в каких случаях она применяется?
3. Приведите типологизацию НКРЯ
4. Какие виды корпусов существуют?

5. Опишите формальные метрики точности работы классификаторов. В чем преимущества и недостатки формальных метрик?

Вторая часть содержит один вопрос, проверяющий ИПК-4, ОПК-2; ПК-4. Ответы на вопросы второй части предполагают решение задач и краткую интерпретацию полученных результатов.

Примерный перечень практических вопросов:

1. Найдите и выпишите 10 примеров перевода/передачи слова «глушь» на английский язык (параллельный корпус НКРЯ).

2. Когда впервые было употреблено слово «толерантность», как менялось его значение на протяжении всех лет его существования

3. С помощью диалектного корпуса найти территории, где употребляются слова:

а) евоный

б) худо

в) пошто

3. Когда впервые было употреблено слово «диверсификация», в какой сфере функционирования данное слово употребляется чаще всего.

4. Используя НКРЯ, проанализируйте левостороннюю сочетаемость слова «стол» с прилагательными, выпишите не менее 10 сочетаний

Результаты экзамена определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Критерии оценки экзамена обусловлены логической демонстрацией приобретенных компетенций в соответствии с текущей программой. Демонстрация предусматривает уверенное использование терминологии, понимание и корректное использование математического аппарата, предусматривает корректность написания кода, его понимание и корректное использование в нем математических методов. Отметка «хорошо» выставляется за счет демонстрации полученных компетенций, владение и понимание кода, теоретических аспектов его применения в практике работы с текстовыми массивами данных допускаются недочеты в понятийном аппарате математики. Отметка «удовлетворительно» позволяет допустить ошибки в разработке кода, но учитывает последовательную логику изложения структуры кода, его интерпретацию, связь теоретических аспектов лингвистики и математики, демонстрация понимания хода обработки текста. Минимальный порог оценки «отлично» составляет 90-100 баллов, хорошо 75-89, удовлетворительно «55-74» ниже 55 – «неудовлетворительно»

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - <https://moodle.tsu.ru/course/view.php?id=12181>

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

в) План семинарских / практических занятий по дисциплине.

Тема 1. Введение в корпусную лингвистику. Введение в корпусную лингвистику

Типология корпусов. Разметка корпуса: лингвистическая и экстралингвистическая разметки. Основные свойства корпуса

Тема 2. Практика по работе в наиболее известных корпусах. Теоретические и практические аспекты использования НКРЯ. Изучение web-интерфейса и функций корпусов: НКРЯ, КРУТ, RLC и пр. Поиск и анализ n-грамм. Конкордансный поиск ЛЕ в корпусе. Семантический поиск

Тема 3. Квантитативные методы исследования. Принципы выбора и обоснование использования квантитативных методов. Абсолютная частота. Относительная частота. Коэффициент Жуйана. Коэффициент логарифмического правдоподобия LL-score. Создание базы данных. Критерий отношения правдоподобия (loglikelihood). Метрики для определения ключевых слов. Самостоятельная работа по теории

Тема 4. Разметка корпуса. Основные принципы и типы разметки корпуса: лингвистическая и экстралингвистическая разметки. Морфологическая разметка корпуса: mystem, AOT, TreeTagger, Marmot. Синтаксическая разметка корпуса: UDPIPE. аннотированная разметка корпуса с помощью ELAN

Тема 5. Разработка корпуса. Принципы разработки корпуса. Парсинг текстов с помощью BootCat. Сборка корпуса в ПО AntConc. Анализ разработанного корпуса

Подготовка к проведению практических работ начинается в начале теоретического изложения изучаемой темы и продолжается по ходу её изучения при освоении материала на занятиях в рамках практических заданий и работе над ним в ходе самостоятельной подготовки дома и в библиотеках. Для качественного выполнения лабораторных работ студентам необходимо:

- 1) повторить теоретический материал по конспекту и учебникам;
- 2) ознакомиться с описанием лабораторной работы;
- 3) в специальной тетради для лабораторных работ записать название и номер работы, перечень необходимого программного обеспечения, подготовить алгоритм или код;

4) выяснить цель работы, четко представить себе поставленную задачу и способы её достижения, продумать ожидаемые результаты опытов;

5) ответить устно или письменно на контрольные вопросы по изучаемой теме или решить ряд задач;

6) изучить порядок выполнения лабораторной работы. Подготовить среду выполнения кода к работе. После проверки правильности алгоритма работы программы преподавателем можно начинать выполнение лабораторной работы.

д) Методические указания по организации самостоятельной работы студентов.

Формы самостоятельной работы студентов разнообразны. Они включают в себя:

- изучение и систематизацию практических и теоретических примеров в рамках выполнения текущих заданий по предмету;
- изучение учебной, научной и методической литературы, материалов периодических изданий с привлечением электронных средств официальной, статистической, периодической и научной информации;
- подготовку докладов и презентаций, написание программного кода и его отладка;
- участие в работе студенческих конференций, комплексных научных исследованиях.

Самостоятельная работа приобщает студентов к научному творчеству, поиску и решению актуальных современных проблем.

Примеры самостоятельной работы студентов:

На основе своего корпуса создайте:

- Частотную матрицу Bag of Words (ключевые слова/части речи/словарь). Веса слов должны быть выражены относительными величинами

- Создайте частотный анализ лексических единиц, используя формулу IPM. Опишите полученный результат, в соответствии со своей гипотезой

- SQL: разработка корпуса. Создайте собственный корпус на базе Python и SQL

Исходный код:

```
CREATE TABLE Lexeme (
    lexicid int not null IDENTITY primary key,
    lex text
);
```

```
CREATE TABLE Lemma (
    lemid int not null IDENTITY primary key,
    lemtype nchar,
    lem text,
    suffix text,
    tag text,
    descr text,
    lexicid int not null references Lexeme(lexicid)
);
```

```
select *
from Lemma
```

```
select *
from Lexeme
```

```
insert into dbo.Lexeme ( lex) values ( 'hkjklj')
```

```
CREATE TABLE Lexeme (
    lexicid int not null identity primary key,
    lex varchar(max)
);
```

```
CREATE TABLE Lemma (
    lemid int not null identity primary key,
    lemtype char,
    lem varchar(max),
    suffix varchar(max),
    tag varchar(max),
    descr varchar(max),
    lexicid int not null references Lexeme(lexicid)
);
```

```
select *
from
dbo.Lexeme
```

```
select *
from
dbo.Lemma
```

```
insert into Dims3.dbo.Lexeme(lex) values ('ssss')
insert into Dims3.dbo.Lemma(lemid, lemtype, lem, suffix, tag, descr, lexicid) values (null,
's', 'd', 'f', 'x', 'h', 10)
```

```
select *  
from  
dbo.Lemma  
where dbo.Lemma.lexid = 1553
```

```
SELECT *  
FROM dbo.Lemma  
JOIN dbo.Lexeme on dbo.Lexeme.lexid = dbo.Lemma.lexid
```

```
SELECT *  
FROM dbo.Lemma  
inner JOIN dbo.Lexeme on dbo.Lexeme.lexid = dbo.Lemma.lexid where  
dbo.Lemma.suffix = 'чик'
```

```
SELECT  
lem,  
lex  
FROM dbo.Lemma  
JOIN dbo.Lexeme on dbo.Lexeme.lexid = dbo.Lemma.lexid where dbo.Lemma.suffix =  
'чик'
```

```
SELECT dbo.Lemma.lem,  
dbo.Lexeme.lex,  
dbo.Lemma.suffix,  
dbo.Lemma.tag,  
dbo.Lemma.descr  
FROM dbo.Lemma  
JOIN dbo.Lexeme ON dbo.Lexeme.lexid = dbo.Lemma.lexid  
WHERE Lemma.suffix = 'чик'
```

```
select *  
from  
dbo.Lexeme  
where dbo.Lexeme.lex='дом'
```

```
SELECT dbo.Lemma.lemtype, dbo.Lexeme.lex, dbo.Lemma.suffix, dbo.Lemma.tag,  
dbo.Lemma.descr  
FROM dbo.Lemma  
JOIN dbo.Lexeme ON dbo.Lexeme.lexid = Lemma.lexid  
WHERE dbo.Lemma.lem = 'домик'
```

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Щипицина Л.Ю. Информационные технологии в лингвистике. Учебное пособие. М., 2013

– Копотев М. Введение в корпусную лингвистику. Учебное пособие. – Прага, 2014.

Н.А. Мишанкина. Базы данных в лингвистических исследованиях. Вопросы лексикографии, Выпуск № 1 (3), 2013 – С.25-33

– Захаров В.П. Корпусная лингвистика: учебник для студентов направления "Лингвистика" / В. П. Захаров, С. Ю. Богданова ; Санкт-Петербургский гос. ун-т. СПб., 2013. 48 с.

– Николаев И.С. / Прикладная и компьютерная лингвистика. Изд. 2 URSS. 2017. 320 с. ISBN 978-5-9710-4633-2.

б) дополнительная литература:

– Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник. СПбГУ РИО , 2013.

– Баранов А.Н. Введение в прикладную лингвистику. – М., 2013.

– Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. Пособие для студентов линг.фак.вузов. – М.: Изд.центр «Академия», 2004. – 208 с.

– Андрущенко В.М. Концепция и архитектура машинного фонда русского языка / Отв. ред. А.П. Ершов. М., 1989.

– Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. Пособие для студентов линг.фак.вузов. – М.: Изд.центр «Академия», 2006. – 304 с.

– Мишанкина Н.А. Основные направления прикладной лингвистики. Учебно-методический комплекс / Н.А. Мишанкина, ТГУ. - Томск, 2010. Режим доступа: <http://edu.tsu.ru/eor/resource/192/tpl/index.html>– ...

в) ресурсы сети Интернет:

– НКРЯ <https://ruscorpora.ru/>

– <https://yandex.ru/dev/mystem/>

– BulTreeBank Bulgarian Treebank

– URL: <http://www.bultreebank.org> CDT Croatian Dependency Treebank

– URL: http://hobs.ffzg.hr/default_en.html CLR Croatian Language Repository

– URL: <http://riznica.ihjj.hr> FIDA Корпус словенского языка FIDA

– URL: <http://www.fida.net> FidaPLUS Корпус словенского языка FidaPLUS

– URL: <http://www.fidaplus.net> PELCRA Polish and English Language Corpora for Research and Applications

– URL: <http://korpus.ia.uni.lodz.pl> Prague Dependency Treebank

–URL: <http://ufal.mff.cuni.cz/pdt> PWN Корпус польского языка издательства PWN

–URL: <http://korpus.pwn.pl/szukaj.php> SDT Slovene Dependency Treebank

–URL: <http://nl.ijs.si/sdt> WWW-Concordance КБТ Корпус боснийских текстов (Осло)

–URL: <http://www.tekstlab.uio.no/Bosnian/Corpus.html> КГТ Корпус газетных текстов русского языка

–URL: <http://www.philol.msu.ru/~lex/corpus> Корпус Института основ информатики Польской академии наук

–URL: <http://korpus.pl> Корпус словенского языка (тематика текстов—«Связи с общественностью»)

–URL: <http://www.korp.fdv>. uni-lj.si Корпус словенского языка Nova beseda

–URL: http://bos.zrc-sazu.si/a_beseda.html Корпус словенского языка

–URL: <http://nl2.ijs.si/index-mono.html>

–Корпусная лингвистика и корпусный подход в обучении иностранному языку
http://ling.ulstu.ru/linguistics/resources/literature/articles/corpus_linguistics_language_teaching/

–URL: <https://ruscorpora.ru/new/corpora-usage.html> Официальная страница центра справки и обучения НКРЯ. [Электронный ресурс].

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.).

– язык программирования R (RStudio) и Python;

– Программа Mystem.

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>

в) профессиональные базы данных:

– Университетская информационная система РОССИЯ – <https://uisrussia.msu.ru/>

– Единая межведомственная информационно-статистическая система (ЕМИСС) –
<https://www.fedstat.ru/>

– Справка ПО и библиотек R-CRAN <https://cran.r-project.org/>

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/>

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.

Лаборатории, оборудованные компьютерами (не ниже i3, RAM 8Gb), проектором

Аудитории для проведения занятий лекционного и семинарского типа индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации в смешенном формате («Актру»).

15. Информация о разработчиках

Степаненко А.А., старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики ФилФ ТГУ