

Министерство образования и науки РФ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Новосибирский национальный исследовательский государственный
университет» (Новосибирский государственный университет, НГУ)
Гуманитарный факультет

Программа рассмотрена
на заседании кафедры
фундаментальной и прикладной
лингвистики
29.08.2014

Зав. кафедрой, проф. М.К. Тимофеева

Утверждаю

декан гуманитарного
факультета, профессор
1.09.2014
Л.Г. Панин

Основная образовательная программа
высшего образования

Направление подготовки
035800 – Фундаментальная и прикладная лингвистика

Квалификация (степень) выпускника –
бакалавр

ПРОГРАММА УЧЕБНОГО КУРСА
«ИНСТРУМЕНТАЛЬНЫЕ И ПРОГРАММНЫЕ МЕТОДЫ ЛИНГВИСТИЧЕСКИХ
ИССЛЕДОВАНИЙ»

(136 часов, 4 з.е.)

1. Наименование дисциплины

ПРОГРАММА УЧЕБНОГО КУРСА «ИНСТРУМЕНТАЛЬНЫЕ И ПРОГРАММНЫЕ МЕТОДЫ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ»

Программа дисциплины «Инструментальные и программные методы лингвистических исследований» составлена в соответствии с требованиями к обязательному минимуму содержания и уровню подготовки дипломированного бакалавра по направлению 035800 «Фундаментальная и прикладная лингвистика» в целях обеспечения реализации учебного процесса в НГУ.

Автор __Леман Анна Яковлевна

2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы.

Цели освоения дисциплины

Дисциплина «Инструментальные и программные методы лингвистических исследований» имеет своей целью: изучение инструментальных методов исследования сенсорных систем человека (зрительной и слуховой), связанных с речью, изучение инструментальных методов исследования работы мозга человека, изучение программного обеспечения, используемого для поддержки лингвистических исследований.

Для достижения поставленной цели выделяются задачи курса:

- изучение теоретических основ исследования сенсорных систем человека (зрительной и слуховой),
- изучение теоретических основ проведения исследования мозга человека (энцефалография, магнитно-резонансная томография),
- изучение редактора онтологий Protégé с проведением практических занятий,
- изучение инструмента для извлечения структурированных данных (фактов) из текста на естественном языке Томита-парсер с проведением практических занятий,
- изучение пакета библиотек и программ для символьной и статистической обработки естественного языка NLTK с проведением практических занятий,
- изучение электронного тезауруса Wordnet с проведением практических занятий.

В результате освоения дисциплины обучающийся должен:

Иметь представление о круге проблем и задачах, связанных с проблематикой курса, о возможных подходах к использованию инструментальных и программных средств для проведения лингвистических исследований, о месте данной дисциплины среди других, об основных областях практического применения полученных знаний.

Знать основы построения онтологий, контекстно-свободных грамматик, электронных тезаурусов и семантических сетей.

Уметь создать онтологию, описывающую предметную область на основе учебных заданий, разработать контекстно-свободную грамматику и применить ее для извлечения фактов Томита-парсером на основе учебных заданий, запрограммировать приложение на языке Python с использованием NLTK и Wordnet.

Перечисленные результаты образования являются основой для формирования следующих общекультурных и общепрофессиональных компетенций:

а) общекультурными (ОК)

- владением культурой мышления, способностью к обобщению, анализу, восприятию информации, постановке цели и выбору путей её достижения ОК-1
- готовностью к кооперации с коллегами, работе в коллективе ОК-3
- стремлением к саморазвитию, повышению своей квалификации и мастерства ОК-6
- способностью применять методы математического анализа и моделирования в профессиональной деятельности ОК-10
- осознанием сущности и значения информации в развитии современного общества; владением основными методами, способами и средствами получения, хранения, переработки информации ОК-11
- владением навыками работы с компьютером как средством управления информацией ОК-12
- способностью работать с информацией в глобальных компьютерных сетях ОК-13

б) профессиональными (ПК):

общепрофессиональными:

- знанием основ математических дисциплин, которые используются при формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур: теории множеств, математического анализа, теории вероятностей и математической статистики, теории информации и кодирования, математической логики, математической теории грамматик ПК 2

в области производственно-практической деятельности и проектной деятельности:

- владением принципами создания представительных текстовых массивов, корпусов текстов, корпусов звучащей речи, мультимодальных корпусов, электронных словарей разных типов, лингвистических баз данных и умением пользоваться этими ресурсами ПК-15
- знанием основ математических дисциплин, которые используются при формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур: теории множеств, математического анализа, теории вероятностей и математической статистики, теории информации и кодирования, математической ПК 2

- умением пользоваться лингвистически ориентированными программными продуктами ПК-16
- способностью применять знания на практике ПК-25
- способностью работать в междисциплинарной команде ПК-26
- способностью общаться с экспертами в других областях знаний ПК-27
- умением видеть междисциплинарные связи изучаемых дисциплин и пониманием их значения для будущей профессиональной деятельности ПК-28

3. Место дисциплины в структуре образовательной программы.

Курс «Инструментальные и программные методы лингвистических исследований» является частью блока БЗ «Профессиональный цикл» (региональный компонент) и используется в других курсах этого раздела.

Программа дисциплины составлена с учетом связей и соотношения учебных дисциплин, преподаваемых на отделении фундаментальной и прикладной лингвистики гуманитарного факультета НГУ.

Преподавание дисциплины базируется на знаниях и умениях, полученных при изучении курсов «Информатика и основы программирования», «Математическая логика», «Дискретные математические модели».

Освоение дисциплины необходимо при специализации в области компьютерных методов фундаментальной и прикладной лингвистики.

4. Объем дисциплины в зачетных единицах с указанием количества академических, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся.

Общая трудоемкость дисциплины составляет 4 зачетных единицы, 136 часов. Из них на контактную работу с преподавателем 68 часов (лекции – 26 часов, семинары и практические занятия – 42 часа), на самостоятельную работу студентов – 68 часов. Занятий в интерактивной форме – 36 часов.

Вид учебной работы	Всего часов	Семестры	
		1	2
Общая трудоемкость дисциплины	136	46	90
Аудиторные занятия, в том числе:			
Лекции	26	10	16
Семинары	28	26	2
Практическая работа	14		14
Самостоятельная работа, в том числе:			
Курсовой проект			
Реферат	10	10	
Расчетные работы			
Другие виды самостоятельной работы			
Разработка самостоятельных проектов с использованием изучаемых технологий	58		58
Вид промежуточного контроля			

5. Содержание дисциплины “Инструментальные и программные методы лингвистических исследований”, структурированное по темам с указанием отведенного на них количества астрономических часов и видов учебных занятий

Новизна курса. В связи с тем, что курс относится к общепрофессиональным дисциплинам, он в первую очередь направлен на изложение фундаментальных знаний по дисциплине. К ним относятся: теория построения онтологий, теория контекстно-свободных грамматик. Новизна курса проявляется в ознакомлении студентов с современными методами и разработками в области лингвистических исследований.

Тематический план курса (распределение часов по видам учебной работы).

№ п/п	Наименование тем и разделов	ВСЕГО (часов)	Аудиторные занятия (часов), в том числе			Самостоятельная работа (часов)
			Лекции	Семинары	Лаб. работы	
1.	Общие подходы к использованию инструментальных и программных средств лингвистических исследований и примеры их использования.	12	2	10		2
2.	Исследование работы мозга и выявление речевых нарушений с использованием МРТ и функционального МРТ	6	2	4		2
3.	Исследование работы мозга и выявление речевых нарушений с использованием ЭЭГ	6	2	4		2
4.	Исследование слуховой системы человека, выявление нарушений, влияющих на речь.	6	2	4		2
5.	Исследование зрительной системы человека, выявление нарушений, влияющих на речь.	6	2	4		2
	Итого за первый семестр:	36	10	26		10
						Дифференцированный зачет
6.	Современные программные средства и технологии, используемые в лингвистических	4	2	2		4

	исследованиях.					
7.	Контекстно-свободные грамматики (КСГ). Использование КСГ для конфигурирования Томита-парсера. Извлечение фактов с помощью Томита-парсера	8	4		4	20
8.	Онтологии. Создание онтологий с использованием Protégé	8	4		4	20
9.	Тезаурусы. Электронный тезаурус Wordnet	4	2		2	4
10.	Написание программ на языке Python для анализа текстов с использованием набора библиотек NLTK.	8	4		4	10
	Итого за второй семестр:	32	16	2	14	58
	ИТОГО:	68	26	28	14	68
						Дифференцированный зачет

6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

Методические рекомендации для самостоятельной работы с литературой и Интернет-ресурсами (перечень вопросов для самостоятельной проработки для каждой темы и темы для подготовки рефератов).

Тема 1. Общие подходы к использованию инструментальных и программных средств лингвистических исследований и примеры их использования

История определения авторства романа «Тихий Дон». Анализ речи душевнобольных. Изучение детской речи. Корпусы звучащей речи. Афазии. Дислексия. Международная конференция по компьютерной лингвистике «Диалог». Сервисы автоматической генерации текстов. Сервисы анализа тональности текстов.

Тема 2. Исследование работы мозга и выявление речевых нарушений с использованием МРТ и функционального МРТ

Речевые патологии на МРТ и ФМРТ. ФМРТ и детектор лжи. Пластичность мозга.

Тема 3. Исследование работы мозга и выявление речевых нарушений с использованием ЭЭГ

История ЭЭГ. Отличия ЭЭГ животных и человека. Связь речевых нарушений и эпилептической активности на ЭЭГ.

Тема 4. Исследование слуховой системы человека, выявление нарушений, влияющих на речь

Речь после кохлеарной имплантации. Особенности письменной речи у глухих людей. Язык жестов. Пиктограммы в коммуникации.

Тема 5. Исследование зрительной системы человека, выявление нарушений, влияющих на речь

Дислексия. Обучение речи слепоглухонемых.

Тема 6. Современные программные средства и технологии, используемые в лингвистических исследованиях.

Big data. N-граммы. POS-теггеры. Использование анализаторов текста для определения плагиата. Использование анализаторов текста для поиска преступников и террористов. Генератор текстов SCIGen. Генерация текстов с помощью цепей Маркова.

Тема 7. Контекстно-свободные грамматики (КСГ). Использование КСГ для конфигурирования Томита-парсера. Извлечение фактов с помощью Томита-парсера

Извлечение фактов при помощи шаблонов. Другие системы извлечения именованных сущностей (фактов)

Тема 8. Онтологии. Создание онтологий с использованием Protégé

Редактор онтологий Fluent Editor. Differential Ontology Editor. Включение онтологий в разрабатываемое ПО. Онтологии в корпоративных системах.

Тема 9. Тезаурусы. Электронный тезаурус Wordnet

Rusnet и его аналоги. Wordnet для китайского языка.

Тема 10. Написание программ на языке Python для анализа текстов с использованием набора библиотек NLTK.

Стемминг Портера. Библиотеки анализа текстов для языка Perl.

Виды внеаудиторной самостоятельной работы студентов по дисциплине “Инструментальные и программные методы лингвистических исследований”

- 1) работа над лекционным материалом;
- 2) работа с литературой;
- 3) работа с Интернет-источниками;
- 4) написание рефератов;
- 5) практическая работа с изучаемыми программными средствами;
- 6) подготовка к зачету.

7. Фонд оценочных средств для промежуточной аттестации обучающихся по дисциплине.

Примерный перечень вопросов к дифференцированному зачету за первый семестр.

История применения инструментальных и программных методов в лингвистических исследованиях. Расположение и специализация речевых зон мозга. МРТ в лингвистических исследованиях. ФМРТ в лингвистических исследованиях. ЭЭГ в лингвистических исследованиях. Строение слуховой системы. Аудиометрия и исследование слуха камертонами. Строение зрительной системы. Методы исследования зрительной системы.

Примерный перечень вопросов к дифференцированному зачету за второй семестр.

Онтологии предметной области. Программы для построения онтологий. Использование Protégé для построения онтологий. Контекстно-свободные грамматики. Программы для извлечения фактов из текста. Применение Томита-парсера для извлечения фактов.

Использование набора библиотек NLTK для программ анализа текстов. Электронные тезаурусы. Электронный тезаурус Wordnet. Электронные тезаурусы русского языка.

8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины.

Основная литература

1. Добров Б. В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. — М.: Бином. Лаборатория знаний, 2009
2. Ахо А., Ульман Дж. Том 1. Теория синтаксического анализа, перевода и компиляции. Синтаксический анализ — М.: Мир, 1978

Дополнительная литература

1. Natural Language Processing with Python. — O'Reilly Media, 2009

10. Методические указания для обучающихся по дисциплине.

Для закрепления теоретического материала и приобретения практических навыков предусмотрено выполнение лабораторных работ на компьютерах.

В процессе выполнения лабораторных работ студенты приобретают навыки разработки онтологий предметной области, конфигурирования Томита-парсера, написания программ для анализа текстов, использующих NLTK. В процессе самостоятельной работы студенты подготавливают рефераты по изучаемым темам и разрабатывают учебные проекты: онтологии предметной области, программы для анализа текстов.

Формами контроля являются: контрольные вопросы студентам после изучения темы, контрольная неделя в середине семестра. Итоговой формой контроля за каждый семестр является дифференцированный зачет.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем.

Protégé, NLTK, Wordnet, Томита, Python

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.

ПК, медиапроектор, мультимедийные презентации