**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего профессионального образования
"Национальный исследовательский университет
"Высшая школа экономики"**

Факультет филологии

**Программа дисциплины** Компьютерные инструменты лингвистических исследований

для направления/ специальности 035800.62 «Фундаментальная и прикладная лингвистика» подготовки бакалавров

Авторы программы:

Толдова С. Ю., к.ф.н., stoldova@hse.ru

Мороз Г. А., agricolamz@gmal.com

Кутузов А. Б. akutuzov72@gmail.com

Одобрена на заседании кафедры [Введите название кафедры] «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 20 г

Зав. кафедрой [Введите И.О. Фамилия]

Рекомендована секцией УМС [Введите название секции УМС] «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 20 г

Председатель [Введите И.О. Фамилия]

Утверждена УС факультета [Введите название факультета] «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_20 г.

Ученый секретарь [Введите И.О. Фамилия] \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ [подпись]

Москва, 2014

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения кафедры-разработчика программы.*

# Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и бакалавров направления подготовки 035800.62 «Фундаментальная и прикладная лингвистика» факультета филологии.

Программа разработана в соответствии с:

* Образовательным стандартом государственного образовательного бюджетного учреждения высшего профессионального образования Высшей школы экономики, в отношении которого установлена категория «национальный исследовательский университет» (ГОБУ ВПО НИУ-ВШЭ) протокол от 02.07.2010
* Образовательной программой направления «Фундаментальная и прикладная лингвистика» подготовки бакалавра;
* Рабочим учебным планом НФ НИУ-ВШЭ на 2014/2015 по направлению подготовки «Фундаментальная и прикладная лингвистика», утвержденным в 2014 году.

# Цели освоения дисциплины

Целями освоения дисциплины «Компьютерные инструменты лингвистических исследований» часть 2 являются знакомство с основными компьютерными инструментами и ресурсами, применяемыми в лингвистических исследованиях. Курс закладывает теоретическую и практическую базу для использования различного инструментария для компьютеризации сбора, обработки и анализа лингвистического материала, а также для представления данных и результатов исследований в виде компьютерных ресурсов.

# Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

*Знать*

* базовые принципы работы с лингвистическими корпусами и ресурсами;
* основные типы и языки запросов к корпусам для поиска материала в соответствии с различными типами задач лингвистических исследований;
* основные методы работы с лингвистическим материалом с использованием редактора LaTeX.
* основные методы и инструменты для разметки собственных корпусов;
* методы представления результатов исследования в виде баз данных и доступных в интернете ресурсов.

*Уметь*

* уметь строить сложные запросы к корпусам с использованием специального языка запросов в соответствии с различными лингвистическими задачами, включая исследование конструкций, анализ коллокаций;
* уметь работать с системой LaTeX, оформлять собственные научные тексты (курсовые работы, рефераты, эссе и т. п.);
* работать с различными типами программ обработки текстов: конкордансерами, программами для поиска коллокаций, создания частотных списков и т.п., корпусными менеджерами, программами для документации языков, включающих поморфемную аннотацию текстов и составление словарей;
* создавать собственную схему разметки к корпусам и реализовывать ее в специальных средах для аннотирования корпусов;
* уметь загружать собственные данные и писать простые запросы к базе данных в СУБД MySQL;
* публиковать свои данные на веб-сайте;

*Иметь навыки (приобрести опыт)*

* работы с материалом, собранным с использованием корпусных ресурсов;
* работы с программами первичной обработки текста;
* работы с редактором LaTeX;
* аннотирования собственных корпусов;
* сбора материала с использованием корпусов;
* представления материала в виде баз данных.

В результате освоения дисциплины студент осваивает следующие компетенции:

| Компетенция | Код по ФГОС/ НИУ | Дескрипторы – основные признаки освоения (показатели достижения результата) | Формы и методы обучения, способствующие формированию и развитию компетенции |
| --- | --- | --- | --- |
| Способен планировать научно-исследовательскую деятельность, проводить самостоятельные исследования и получать новые научные результаты в области профессиональной деятельности | ПК-1 | понимает постановку задачи лингвистического исследования с точки зрения использования возможностей соответствующих электронных ресурсов для сбора лингвистического исследования; умеет применять простые компьютерные инструменты для сбора лингвистических данных и их обработки | - чтение специальной литературы-выполнение самостоятельных заданий по поиску в корпусах- анализ полученных данных с использованием специальных программ |
| работать с электронными словарями и другими электронными ресурсами для решения лингвистических задач | ПК-17 | знает основные лингвистические ресурсы; умеет их использовать  | практические занятия по использованию языковых ресурсов и компьютерных инструментов лингвистического анализа |

# Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу профессиональных дисциплин, обязательных для изучения.

Изучение данной дисциплины базируется на следующих дисциплинах:

* Теория языка (программы подготовки бакалавра)

Желателен опыт научно-исследовательской работы, проводимой в рамках различных НИС лингвистического направления.

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

* владеть базовыми представлениями о грамматических категориях и анализе языковых единиц
* владеть базовыми навыками работы с компьютером

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

* корпусная лингвистика, компьютерная лингвистика, научно-исследовательские семинары по различным лингвистическим задачам

# Тематический план учебной дисциплины

Часть 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Название раздела | Всего часов  | Аудиторные часы | Самостоя­тельная работа |
| Лекции | Семинары | Практические занятия |
| 1 | Работа с корпусными менеджерами, использующими специальный язык запросов |  |  |  | 8 | 10 |
| 2 | Инструменты аннотации корпусов | 38 |  |  | 8 | 10 |
|  | Инструменты представления данных средствами LaTeX |  |  |  | 16 | 10 |
| 3 | Инструменты анализа и представления данных | 40 |  |  | 8 | 20 |
|  | Итого | 90 |  |  | 40 | 50 |

# Формы контроля знаний студентов

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | 1 | 2 | 3 |  |  |
| Тип контроля | Форма контроля | 2 год | Параметры \*\* |
| Итоговый | Экзамен |  |  | 10 |  | Экзамен в виде выполнения итогового проектного задания (письменная и устная часть)Поиск с использованием специального языка запросовРазметка собственного корпусаРедактирование текста и представление лингвистических данных в LaTeXОдно из типов заданий на выбор:Базы данных. Разработки БД для представления данных собственных исследований. СУБД MySQLРабота с параллельными корпусамиОбработка аудио данных в специальных редакторах работы с аудиоданными |

## Критерии оценки знаний, навыков

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

Домашнее задание по характеристике корпуса является групповым проектом

# Содержание дисциплины

Второй год

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Название | лекции | практические занятия | Литература или сетевые ресурсы по разделу |
| Раздел 1. Лингвистические ресурсы |
|  | Корпусные менеджеры |  | 2 | <http://cwb.sourceforge.net/> <http://corpus.leeds.ac.uk/list.html> <http://cwb.sourceforge.net/files/CQP_Tutorial/>  |
|  | Поиск в корпусе с использованием специального языка запросов CQP |  | 2 |
|  | Поиск коллокаций |  | 2 |
|  | Использование CQP и поиска коллокаций, средств работы с таблицами Excel и конкордансеров для сбора и обработки данных |  | 2 |
| Раздел Аннотирование корпусов |
|  | Brat - веб-технологии глубокой разметки корпусов - знакомство с Brat  |  | 2 | <http://brat.nlplab.org> – специализированная среда для аннотирования корпусов<http://brat.nlplab.org/manual.html> - документация по работе с bratPontus Stenetorp, Sampo Pyysalo, Goran Topic, Sophia Ananiadou, Akiko Aizawa. Normalisation with the brat rapid annotation tool. In Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine, Zürich, Switzerland, September 2012.. <http://www.aclweb.org/anthology/E12-2021>  |
|  | Создание собственной схемы разметки в среде Brat |  | 4 |
|  | Конвертация данных из одной схемы разметки в другую |  | 2 |
|

|  |
| --- |
| Раздел Инструменты для анализа и обработки текста |

 |
|  | Введение. Что такое издательская система LaTeX? Установка системы и основные принципы работы.  |  | 2 | Львовский С. М. «Набор и верстка в системе LaTeX». 3-е издание, исправленное и дополненное. Электронная версия: <http://tex.bog.msu.ru/Lvovskij.pdf>  |
|  | Верстка таблиц и графиков. Оформление абзацев и текста в целом |  | 2 |  |
|  | Работа с библиографией |  | 2 |  |
|  | Набор знаков IPA, встройка разных систем письма в LaTeX, организация языковых примеров и глоссирование, оформление кода языков программирования |  | 4 |  |
|  | Интеграция R в LaTeX, представления вокалических систем, рисование просодических и синтаксических деревьев |  | 4 |  |
|  | Средства, позволяющие делать презентации и постеры в LaTeX |  | 2 |  |
| Раздел Инструменты для анализа и представления данных |
|  | Работа с базами данных: СУБД MySQL  |  | 4 | Практикум по MySQL <http://hsemysql.wikispaces.com/home> Практикум <http://hsemysql.wikispaces.com/php1> |
|  | Работа с параллельными корпусами |  | 4 |  |
|  | Работа с редакторами обработки звука и создания аудио корпусов |  | 4 | <http://www.fon.hum.uva.nl/praat/> <http://savethevowels.org/praat/> <http://www.mpi.nl/corpus/html/elan/><http://www.mpi.nl/corpus/manuals/manual-elan_ug.pdf>  |

# Образовательные технологии

Рекомендуемые образовательныетехнологии включают лекции, практические занятия, самостоятельную работу студентов (выполнение практических домашних заданий с использованием специализированного компьютерного инструментария).

При проведении занятий рекомендуется использование интерактивных форм занятий (проектных методик, разбор конкретных ситуаций, включение в лекционный курс интерактивного общения с аудиторией, презентаций, контрольных вопросов на понимание) в сочетании с внеаудиторной работой. Удельный вес занятий, проводимых в интерактивных формах, должен составлять не менее 40 % аудиторных занятий.

# Оценочные средства для текущего контроля и аттестации студента

## Тематика заданий текущего контроля

Задания текущего контроля включают задания разной степени сложности: тренировочные задания для закрепления навыков, полученных студентами в ходе практических занятий, а также итоговые задания по каждому из разделов.

Примерные вопросы/ задания для текущей проверки усвоения навыков:

1. Раздел 1. Лингвистические ресурсы

Выполнить запросы к одному из корпусов русского языка, представленных на сайте <http://corpus.leeds.ac.uk/ruscorpora.html>: найти примеры со словами некоторой словообразовательной модели (например, глаголы с приставкой *пере* и суффиксом *–ся*); построить запрос, чтобы получить как можно более точное множество примеров конструкции глагол + прямое дополнение; с использованием опции поиск коллокаций найдите коллокации некоторой лексемы, сравните их со словарем.

1. Раздел 2. Аннотирование собственных корпусов

Выполнить разметку фрагмента корпуса в среде Brat. Добавить в схему разметки новый признак. Выполнить импорт новых текстов для разметки. Произвести сравнение разметки.

1. Раздел 3. Инструменты для анализа и обработки текста. Издательская система LaTeX

Воспроизведите средствами LaTeX графическое изображение фрагмента научного текста.

Примеры итоговых заданий для разных разделов программы:

1. Используя Excel, конкордансер AntConc, поиск в корпусах на русском языке на сайте <http://corpus.leeds.ac.uk/ruscorpora.html> проведите мини исследование одной из конструкций.

# Разметка исследовательского корпуса

Задача: разметить свой исследовательский корпус по соответствующим параметрам, провести анализ результатов: получить подвыборки корпуса по определенным значениям признаков, провести статистический анализ данных.

В результате выполнения задания необходимо получить:

1. Размеченный по созданной Вами схеме корпус текстов (не менее 300 предложений)
2. Схему разметки
3. Подвыборки из корпуса по одному признаку, по комбинации признаков
4. Привести глобальную статистику по всему размеченному корпусу, привести примеры статистических данных по одному из значений признака, локальную по одному из признаков, сравнительную (результаты статистического анализа прокомментировать).

Для выполнения задания:

1. Создайте свой исследовательский корпус по интересующей Вас проблеме, либо по одной из проблем предложенных в приложении. Корпус должен включать не менее 300единиц разметки**.**
2. Вы можете воспользоваться корпусом, полученным в результате выполнения задания по поиску в корпусе.
3. Можете взять свой корпус, созданный по любым другим источникам, которыми Вы пользуетесь для написания курсовой
4. Подготовьте корпус к работе в системе
5. Создайте схему разметки в которую войдет не менее 3-4-х параметров:

из них не менее 2-х независимых систем признаков (например, «структура возвратного местоимения» (составное vs. простое) и его синтаксическая позиция), не менее 2-х иерархически организованных систем (т.е. областью определения некоторого признака Y является только подмножество единиц разметки (предложений), которым приписано некоторое конкретное значение A признака X; например, по признаку «предлог» (какой предлог используется с возвратным местоимением) размечаются только предложения, в которых признак «Синтаксическая позиция возвратного местоимения» имеет значение «предложная группа»)

1. Разметьте Ваш корпус
2. Задайте соответствующие фильтры результаты сохраните в отдельных файлах
3. Проведите статистический анализ, результаты сохраните в соответствующих файлах. В отдельном файле кратко прокомментируйте результаты статистической обработки размеченного корпуса
4. Домашнее задание — самостоятельная научная работа студента (эссе, реферат, домашнее задание, курсовая работа и т. п.), выполненная по всем правилам самостоятельно средствами LaTeX с демонстрацией навыков, полученных при освоении разделов 1—7 программы (или разделов 1—8, если работа выполнена в форме презентации). Допускается выполнение и сдача работы по другой дисциплине, выполненной средствами LaTeX, или выполнение работы специально для данного курса.
5. Создайте таблицы по данным курсовой, загрузите в базу данных с использованием СУБД MySQL, постройте SQL запросы к этим таблицам

\* Опубликуйте на сайте таблицу с запросами к Вашей базе данных

## Вопросы для оценки качества освоения дисциплины

Какие корпусные менеджеры Вы знаете? Что представляет собой корпусной менеджер?

Какие коллокационные меры Вы знаете? Что такое коллокации? Зачем нужны коллокационные меры при работе с корпусными данными?

Что такое глубокое аннотирование корпуса? Какие типы аннотирования Вы знаете?

# Порядок формирования оценок по дисциплине

Преподаватель оценивает проектную и индивидуальную самостоятельную работу студентов, работу на семинарских и практических занятиях – *Оаудиторная*.

Преподаватель оценивает самостоятельную работу студентов: оценивается правильностть выполнения самостоятельных домашних работ, отчетов и творческих заданий. Оценки за самостоятельную работу студента преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале за самостоятельную работу определяется перед промежуточным или итоговым контролем – *Осам. работа*.

Результирующая оценка за промежуточный (итоговый) контроль в форме экзамена выставляется по следующей формуле, где *Оэкзамен* – оценка за работу непосредственно на экзамене:

 *Оитоговый = k1·Оэкзамен + k2·Осам. работа + k3·Оаудиторная*

Способ округления накопленной оценки промежуточного (итогового) контроля в форме экзамена: в пользу студента.

При этом удельный вес форм контроля распределяется следующим образом:

*k1 = 0,4*

*k2 = 0,55*

*k3 = 0,05*

На пересдаче студенту предоставляется возможность получить дополнительный балл для компенсации оценки за текущий контроль.

# Учебно-методическое и информационное обеспечение дисциплины

##### Основная литература

Львовский С. М. «Набор и верстка в системе LaTeX». 3-е издание, исправленное и дополненное. Электронная версия: <http://tex.bog.msu.ru/Lvovskij.pdf>

The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial. 2. Basic CQP features. <http://cwb.sourceforge.net/files/CQP_Tutorial/node6.html>

<http://brat.nlplab.org/manual.html> - документация по работе с brat

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Sophia Ananiadou, Akiko Aizawa. Normalisation with the brat rapid annotation tool. In Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine, Zürich, Switzerland, September 2012.. <http://www.aclweb.org/anthology/E12-2021>

##### Дополнительная литература

<http://studiorum.ruscorpora.ru/index.php?option=com_docman&Itemid=111> - примеры корпусных исследований лексики в исторической перспективе

<http://office.microsoft.com/ru-ru/excel-help/>

Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). URL: <http://www.dialog-21.ru/digests/dialog2010/materials/pdf/22.pdf>

Christopher Manning and Hinrich Schütze. [Foundations of Statistical Natural Language Processing](http://www-nlp.stanford.edu/fsnlp/promo/). Chapter 5. Collocations. MIT Press. 1999. URL: <http://nlp.stanford.edu/fsnlp/promo/colloc.pdf> (URL: <http://www.alingva.ru/articles/collocation_rus.pdf> - (русский перевод главы 5))

## Программные средства и Интернет-ресурсы

Для успешного освоения дисциплины студент должен использовать следующие программные средства/ресурсы:

<http://corpus.leeds.ac.uk/ruscorpora.html>

система LaTeX

<http://brat.nlplab.org> – система аннотирования корпусов

PRAAT - http://praat.org/).

Speech Analyzer (http://www-01.sil.org/computIng/sa/)

## Дистанционная поддержка дисциплины

В части курса (MySQL и PHP) используются онлайн-практикумы, содержащие подробные инструкции и обучающие задания.

# Материально-техническое обеспечение дисциплины

Для проведения практических занятий необходимы компьютерные классы, для проведения лекций проектор, экран