**Правительство Российской Федерации**

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

Факультет

**Программа дисциплины** Компьютерные инструменты лингвистических исследований

для направления/ специальности подготовки бакалавров

Авторы программы:

Толдова С.Ю., к.ф.н., toldova@yandex.ru

Ляшевская О.Н., к.ф.н., olesar@gmail.com

Одобрена на заседании кафедры «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 20 г

Зав. кафедрой

Рекомендована секцией УМС «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 20 г

Председатель

Утверждена УС факультета «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_20 г.

Ученый секретарь \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва, 201\_

*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 2012/2013 по направлению подготовки «Фундаментальная и прикладная лингвистика», утвержденным в 2012 году.

# Цели освоения дисциплины

Целями освоения дисциплины «Компьютерные инструменты лингвистических исследований» 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*

* основные типы компьютерных лингвистических ресурсов, используемых для сбора материала исследований;
* базовые принципы работы с лингвистическими корпусами и ресурсами;
* основные типы запросов к корпусам для поиска материала в соответствии с различными типами задач лингвистических исследований;
* основные методы работы с материалом с использованием различных опций Microsoft Excel.
* методы обработки материала с помощью специальных компьютерных инструментов, таких как конкордансеры;
* методы представления результатов исследования в виде баз данных и доступных в интернете ресурсов.

*Уметь*

* работать с простыми средствами обработки текстов: текстовыми редакторами и редакторами электронных таблиц, таких как Excel;
* осуществлять оценку различных типов современных корпусных ресурсов и выбирать ресурсы, подходящие для выполнения тех или иных исследовательских и производственных задач;
* осуществлять поиск в корпусах, в том числе и с использованием специальных языков запросов, в соответствии с исследовательской гипотезой в области грамматики и лексикографических исследований;
* 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;
* 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 текстовыми редакторами и электронными таблицами;
* 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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Название раздела | Всего часов | Аудиторные часы | | | Самостоя­тельная работа |
| Лекции | Семинары | Практические занятия |
| 1 | Лингвистические ресурсы |  | 6 |  | 6 | 18 |
| 2 | Инструменты для анализа и обработки текста |  | 8 |  | 8 | 20 |
| 3 | Инструменты для анализа и представления данных |  | 6 |  | 6 | 20 |

# Формы контроля знаний студентов

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | | | | Параметры \*\* |
| 1 | 2 | 3 |  |
| Текущий  (неделя) | Домашнее задание |  | 4 |  |  | 1. Задание по поиску в Национальном корпусе русского языка 2. Доклад: характеристика корпуса 3. Обработка данных в Excel 4. Обработка данных с помощью конкордансера AntConc и дальнейший анализ данных в Excel |
|  |  | 4 |  | 1. Задание по поиску с использованием языка SQP в одном из корпусов, доступных с сайта университета Лидса 2. Глоссирование текста в среде SIL FieldWorks. Написание правил фонетических чередований в среде SIL FieldWorks 3. Базы данных. Разработки БД для представления данных собственных исследований. СУБД MySQL 4. Разработка и публикация страницы с данными из БД |
| Итоговый | Зачет | 3 |  |  |  | Зачет в виде выполнения итогового зачетного задания |

## 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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Название | лекции | практические занятия | Литература или сетевые ресурсы по разделу |
| Раздел Лингвистические ресурсы | | | | |
| 1. 1 | Типы лингвистических ресурсов. Специальные базы данных, корпуса, лексикографические ресурсы | 2 |  | Плунгян В. А. [Зачем нужен Национальный корпус русского языка? Неформальное введение](http://ruscorpora.ru/sbornik2005/02plu.pdf) // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 6—20  Савчук С. О. [Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции](http://ruscorpora.ru/sbornik2005/05savchuk.pdf) // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 62—88  <http://www.ruscorpora.ru/corpora-parameter.html> - о метаразметке  О. Н. Ляшевская, С. А. Шаров, Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.  Ресурсы и корпуса:  <http://ruscorpora.ru/>  <http://corpus.byu.edu/coca/>  <http://wordnetweb.princeton.edu/perl/webwn>  <http://visuwords.com/>  <http://thesaurus.com/>  <http://dict.ruslang.ru/>  <http://starling.rinet.ru/babel.php?lan=ru>  <http://www.ethnologue.com/> |
| 1. 2 | Корпуса как вид лингвистических данных | 2 | 2 |
| 1. 3 | Особенности поиска в Национальном корпусе русского языка |  | 2 |
| 1. 4. | Характеристики и поисковые возможности основных корпусов | 2 | 2 |
| Раздел Инструменты для анализа и обработки текста | | | | |
|  | Excel: различные текстовые функции и их комбинации, полезные для работы со словарными списками | 2 | 4 | Ресурсы и программы с прилагающейся на сайтах документацией:  <http://office.microsoft.com/ru-ru/excel-help/> |
|  | Работа с собственными корпусами текстов. Инструменты обработки текстов - задачи и требования к конкордансерам, знакомство с AntConc | 2 |  | <http://www.antlab.sci.waseda.ac.jp/software.html>  <http://www.antlab.sci.waseda.ac.jp/software/antconc_guide_by_warren_tang_20110305.pdf> - инструкция по работе с AntConc  Частотный списки для русского языка:  <http://corpus.leeds.ac.uk/serge/frqlist/> |
|  | Специальный язык запросов к корпусам. Корпусной менеджер CQP | 2 |  | <http://corpus.leeds.ac.uk/mocky/msd-ru.html> - спецификация для обозначения грамматических категорий русского языка |
|  | Глоссирование в программе FieldWorks | 2 | 4 | <http://fieldworks.sil.org/download/> |
| Раздел Инструменты для анализа и представления данных | | | | |
|  | Работа с базами данных: СУБД MySQL | 2 | 2 | Практикум по MySQL <http://hsemysql.wikispaces.com/home> |
|  | Публикация собственных данных на сайте. Краткое введение в PHP | 2 | 4 | Практикум <http://hsemysql.wikispaces.com/php1> |
|  | Создание собственного ресурса при работе над курсовой или проектом НИС | 2 |  |  |

# 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

Примерные вопросы/ задания для домашних работ:

1. С использованием различных опций поиска в Национальном корпусе русского языка (НКРЯ) и опций работы свыдачей: найдите примеры со словами, начинающиеся на *псевдо-*, но не примеры с лексемой *пседоним*; проследите исторические изменения в значении и употреблении лексемы (например, лексемы *власть*); найдите предложения, порядок слов в начале которых глагол+подлежащее; ответьте на вопрос: какого цвета бывают эмоции в русском языке.
2. Подготовьте презентацию по одному из корпусов по соответствующей схеме описания корпуса.
3. С использованием Excel обработайте массив примеров из НКРЯ на причастия прошедшего времени; определите, какие глагольные основы встречаются с суффиксом –*ну,* какие без него, какие в обоих вариантах, получите частоту по соответствующим способам образования причастий для каждой из найденных в корпусе основ.
4. Обработайте собственные тексты (объемом не менее 500 тыс слов) с помощью программы AntConc; создайте список наиболее частотных терминов вашего корпуса, сравните его со списком Нового частотного словаря (<http://corpus.leeds.ac.uk/serge/frqlist/>); получите конкорданс, информацию о биграмах, информацию об устойчивых словосочетаниях по собственному корпусу для одной из лексемы
5. Выполните запросы к одному из корпусов русского языка, представленных на сайте <http://corpus.leeds.ac.uk/ruscorpora.html>: найдите примеры со словами некоторой словообразовательной модели (например, глаголы с приставкой *пере* и суффиксом *–ся*); постройте 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
6. Отглоссируйте текст впрограмме FieldWorks, напишите правило морфонологического чередования
7. Создайте таблицы по данным курсовой, загрузите в базу данных с использованием СУБД MySQL, постройте SQL запросы к этим таблицам
8. Опубликуйте на сайте таблицу с запросами к Вашей базе данных

## Вопросы для оценки качества освоения дисциплины

1. Какие типы лингвистических данных вам известны?
2. Какие требования предъявляются к составлению корпусов?
3. Каковы методы оценки частотности слова в корпусе?
4. Какие типы корпусов 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 *Озачет* – оценка за работу непосредственно на экзамене:

*Оитоговый = k1·Озачет + 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

Плунгян В. А. [Зачем нужен Национальный корпус русского языка? Неформальное введение](http://ruscorpora.ru/sbornik2005/02plu.pdf) // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 6—20

Савчук С. О. [Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции](http://ruscorpora.ru/sbornik2005/05savchuk.pdf) // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 62—88

## 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

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

* <http://www.ruscorpora.ru/>
* <http://corpus.leeds.ac.uk/ruscorpora.html>
* средства Microsoft Office
* конкордансер AntConc (<http://www.antlab.sci.waseda.ac.jp/software.html>)
* программные среды SIL FieldWorks (<http://fieldworks.sil.org/download/>)

В программе также используются электронные ресурсы Wordnet, Framenet, база данных "Частотного словаря русского языка (на материалах НКРЯ)", база данных Грамматического словаря русского языка А.А.Зализняка.

## 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