**Правительство Российской Федерации**

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

Факультет филологии

**Программа дисциплины**

«Проектирование лингвистических ресурсов и систем»

для направления 035800.68 «Фундаментальная и прикладная лингвистика» подготовки магистра для магистерской программы «Компьютерная лингвистика»

Авторы программы:

Бонч-Осмоловская А.А., к.ф.н, abonch@gmail.com

Ляшевская О.Н., к.ф.н., olesar@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, 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 студентов направления подготовки/ специальности 035800.68 «Фундаментальная и прикладная лингвистика» обучающихся по магистерской программе «Компьютерная лингвистика» изучающих дисциплину Проектирование лингвистических ресурсов и систем.

Программа разработана в соответствии с:

* Образовательным стандартом федерального государственного автономного образовательного учреждения высшего профессионального образования национального исследовательского университета «Высшая школа экономики», в отношении которого установлена категория «национальный исследовательский университет»
* Учебным планом университета по направлению подготовки 035800.68 «Фундаментальная и прикладная лингвистика» для подготовки магистра для магистерской программы «Компьютерная лингвистика» утвержденным в 2012г.

# Цели освоения дисциплины

Целями освоения дисциплины «Проектирование лингвистических ресурсов и систем» являются формирование навыков по созданию лингвистических компонентов обработки текстов, формирование навыков работы в команде, формулирования и решения поставленных задач, методологически корректного построения экспериментов и исследований.

# Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

*Знать*

* основные принципы разработки и создания лингвистических корпусов и ресурсов;
* преимущества и недостатки формальных моделей, лежащих в основе различных модулей автоматической обработки текста;
* необходимые этапы морфологического анализа и проблемы, возникающие при моделировании каждого из этапов;
* основные алгоритмы, используемые для построения автоматического синтаксического анализа;
* наиболее известные доступные для свободного использования компоненты автоматического анализа, в том числе синтаксические и морфологические парсеры, системы распознавания именованных сущностей (NER);

принципы оценки качества таких систем.

*Уметь*

* проводить оценку качества систем автоматического морфологического, синтаксического и семантического анализа;
* использовать соответствующие модули в различных приложениях;
* участвовать в разработке отдельных модулей в системах автоматического анализа текстов.

*Иметь навыки (приобрести опыт)*

* самостоятельной разработки лингвистических ресурсов

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 рефлексировать (оценивать и перерабатывать) освоенные научные методы\* и способы деятельности | СК- 1 | Умеет применять научные методы на практике, умеет критически анализировать теоретические и практические знания в области лингвистики и смежных наук, | Проектная работа, отчетность по проектной работе |
| Способен предлагать концепции\*, модели, изобретать и апробировать способы и инструменты профессиональной деятельности | СК- 2 | Умеет самостоятельно формализовать подходы к задачам, умеет разработать вспомогательные инструменты для осуществления профессиональных задач | Проектная работа, отчетность по проектной работе |
| Способен к самостоятельному освоению новых методов исследования, изменению научного и научно-производственного профиля своей деятельности | СК- 3 | Ориентируется в современной литературе профессиональной тематике, умеет применить на практике методы, описаываемые в научных работах | Проектная работа, отчетность по проектной работе |
| Способен совершенствовать и развивать свой интеллектуальный и культурный уровень, строить траекторию профессионального развития и карьеры | СК- 4 | Владеет навыками стратегического планирования проектной задачи, использует потенциал привлечения ресурсов из смежных дисциплин | Проектная работа, отчетность по проектной работе |
| Способен принимать управленческие решения, оценивать их возможные последствия и нести за них ответственность | СК- 5 | Умеет самостоятельно анализировать достоинства и недостатки подготовленных отчетов, владеет техникой презентации результатов проделанной работы | Проектная работа, отчетность по проектной работе |
| Способен анализировать, верифицировать, оценивать полноту информации в ходе профессиональной деятельности, при необходимости восполнять и синтезировать недостающую информацию и работать в условиях неопределенности | СК- 6 | Активно участвует в критическом обусждении проектов, владеет техниками экспертных оценок лингвистических данных, умеет восполнять недостающие данные с помощью моделирования | Проектная работа, отчетность по проектной работе |
| Способен организовать многостороннюю (в том числе межкультурную) коммуникацию и управлять ею | СК- 7 | Владеет навыками организации проектной работы, имеет опыт консультаций по проекту с внешними экспертами | Проектная работа, отчетность по проектной работе |
| Способен вести профессиональную, в том числе научно-исследовательскую деятельность в международной среде | СК- 8 | Имеет навыки представления отчетов о проектной работе на иностранном языке | Проектная работа, отчетность по проектной работе |
| Способен планировать научно-исследовательскую деятельность, проводить самостоятельные исследования и получать новые научные результаты в области профессиональной деятельности | ПК-1 | понимает постановку задачи в области компьютерной лингвистики, использует современные методы тестирования качества, применяет современные подходы к решению задач в области компьютерной лингвистики | Проектная работа, отчетность по проектной работе |
| Способен выбирать оптимальные теоретические подходы и методы решения конкретных задач вобласти лингвистики и междисциплинарных исследований на стыке различных областей научного знания | ПК-2 | Умеет применять научные методы на практике, умеет критически анализировать теоретические и практические знания в области лингвистики и смежных наук, | Проектная работа, отчетность по проектной работе |
| Способен произвести углубленный анализ проблем, постановку и обоснование проектных задач, связанных с лингвистической проблематикой, моделированием и разработкой лингвистических компонент в междисциплинарных проектах | ПК-5 | понимает постановку задачи в области компьютерной лингвистики, использует современные методы тестирования качества, применяет современные подходы к решению задач в области компьютерной лингвистики, привлекает знания из смежных дисциплин | Проектная работа, отчетность по проектной работе |
| Способен самостоятельно разработать методический инструментарий для осуществления исследовательской и проектной деятельности в области фундаментальной и прикладной лингвистики | ПК-6 | Владеем навыками разработки вспомогательных инструментариев для реализации проектных задач | Проектная работа, отчетность по проектной работе |
| Способен анализировать языковые данные с применением качественных и количественных методов, различных инструментальных средств | ПК-10 | Владеет техникой применения количественного и качественного анализа языковых данных, умеет применять различные инструментальные среды и средства | Проектная работа, отчетность по проектной работе |
| Способен разрабатывать и совершенствовать электронные языковые ресурсы, лингвистические компоненты информационных систем, лингвстические модули компьютерных систем обучения | ПК-17 | знает основные лингвистические ресурсы, владеет цепочкой лингвистической обработкой текста, умеет использовать среды разработки лингвистических систем | проектные решения, использование сред разработки лингвистических технологий |
| Способен разрабатывать и совершенствовать электронные языковые ресурсы, лингвистические компоненты информационных систем, лингвистические модули компьютерных систем обучения | ПК-18 | Имеет опыт проектного решения задач связанных с разработкой лингвистических электронных ресурсов, лингвистических компонентов информационных систем, лингвистических модулей систем обучения | проектные решения, использование сред разработки лингвистических технологий |
| Способен проводить квалифицированный анализ различных типов устного и письменного дискурса в производственно-практических целях | ПК-19 | Умеет классифицировать корпуса текстов по микрожанрам дискурса, выделять ключевые слова и термины, анализировать формальные представления неструктурированных данных | проектные решения, использование сред разработки лингвистических технологий |
| Способен разрабатывать методики и алгоритмы дигитализации гуманитарного наследия, в том числе специальные языковые разметки корпусов | ПК-20 | Знает принципы разметки текстов, владеет основными представлениями о семантической разметки гуманитарного наследия, владеет методами анализа xml-документов | проектные решения, использование сред разработки лингвистических технологий |

# Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу профессиональных дисциплин, обязательных для изучения

Изучение данной дисциплины базируется на следующих дисциплинах:

* Введение в лингвистику (адаптационный курс) или курс по теории языка программы подготовки бакалавра
* Введение в математику (адаптационный курс) или курс по дискретной математики программы подготовки бакалавра
* Введение в программирование (адаптационный курс) или начальный курс по программированию программы подготовки бакалавра
* Иностранный язык

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

* Владеть базовыми представлениями о грамматических категориях и анализе языковых единиц
* Владеть базовыми знаниями в области теории вероятностей и статистики
* Уметь читать научные работы и технические описания на английском языке

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

* Машинный перевод, корпусная лингвистика, онтологии и семантические технологии

# Тематический план учебной дисциплины

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Название раздела | Всего часов | Аудиторные часы | | | Самостоя­тельная работа |
| Лекции | Семинары | Практические занятия |
| 1 | Введение в проблематику прикладной лингвистики | 72 |  | 26 |  | 46 |
|  | Проектная работа | 386 |  | 126 |  | 260 |
|  | Программирование по проекту | 190 |  | 48 |  | 142 |
|  | Итого | 648 |  | 200 |  | 448 |

# Формы контроля знаний студентов

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | | | | 2 год | | Параметры \*\* |
| 1 | 2 | 3 | 4 | 1 | 2 |
| Текущий  (неделя) | Домашняя работа |  | 6 |  |  |  |  | Подготовка плана проекта. Подготовка технического задания и планирование основных временных рамок |
| Домашняя работа над проектом |  |  | 6 |  |  |  | Написание еженедельных отчетов о ходе проекта |
|  | Коллоквиум |  |  |  |  | 3 | 2 | Представление промежуточных результатов проекта и коллективное обсуждение. Участие в коллективном обсуждении других проектов |
|  | Коллоквиум |  |  |  |  |  |  | Представление промежуточных результатов проекта и коллективное обсуждение. Участие в коллективном обсуждении других проектов |
|  | Коллоквиум |  |  |  |  |  |  | Представление промежуточных результатов проекта и коллективное обсуждение. Участие в коллективном обсуждении других проектов |
|  | Коллоквиум |  |  |  |  |  |  | Представление промежуточных результатов проекта и коллективное обсуждение. Участие в коллективном обсуждении других проектов |
|  | Коллоквиум |  |  |  |  |  |  | Представление промежуточных результатов проекта и коллективное обсуждение. Участие в коллективном обсуждении других проектов |
| Итоговый | Финальная презентация |  |  |  | 1 |  | 1 | Открытая презентация проекта |

## Критерии оценки знаний, навыков

По самостоятельной работе и презентации работы над проектом выставляется интегрированная оценка по десятибалльной шкале. Финальная оценка проекта выставляется по 10 бальной шкале. Оценивается активность и участие в семинарах, выполнение домашних заданий, умение использовать теоретические знания из других курсов для решения проектных задач.

Материалы проектных работ публикуются в открытом доступе в Интернет на одной из общедоступных платформ. Презентации работы над проектом проходят в регулярном режиме, не реже, чем два раза в модуль.

## Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на семинарских и практических занятиях:. Оценки за работу на семинарских и практических занятиях преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале за работу на семинарских и практических занятиях определяется перед промежуточным или итоговым контролем - *Оаудиторнаы я*.

Преподаватель оценивает самостоятельную работу студентов. 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*.

Преподаватель оценивает проектную работу студентов, проекты выполняются группой студентов. Накопленная оценка по 10-ти балльной шкале за работу на семинарских и практических занятиях определяется перед промежуточным или итоговым контролем – *Опроектная*

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

*Опромежуточный = k1·Оэкзамен + k2·Опроектная + k3·Осам. работа + k4·Оаудиторная*

При этом удельный вес форм контроля распределяется следующим образом:

*k1 = 0,2*

*k2 = 0,4*

*k3 = 0,2*

*k4 = 0,2*

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

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

Итоговая оценка формируется как сумма двух промежуточных оценок *Опромежуточный*  и выставляется по следующей формуле:

*Оитоговая = k1·Опромежуточная\_1 + k2·Опромежуточная* \_2

При этом удельный вес форм контроля распределяется следующим образом:

*k1 = 0,6*

*k2 = 0,4*

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

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

# Содержание дисциплины

|  |  |  |
| --- | --- | --- |
|  | Название | практические занятия |
| Раздел  Введение в проблематику компьютерной лингвистики | | |
| 1 | История компьютерной лингвистики | 6 |
| 2 | Уровни обработки лингвистической информации | 6 |
| 3 | Ключевые темы современной компьютерной лингвистики | 6 |
| 4. | Современной состояние российской компьютерной лингвистики | 4 |
| 5. | Тенденции развития компьютерной лингвистики | 4 |
| Раздел  Проектная работа над лингвистическими ресурсами | | |
| 1. | Предварительный анализ, постановка задач по каждому проекту | 24 |
| 2. | Анализ выполненной работы по проекту, обработка данных, обсуждение методов | 84 |
| 3. | Программирование в рамках проекта | 48 |
| 3. | Итоговая презентация проекта | 18 |

Литература по разделам:

Раздел «Введение в проблематику компьютерной лингвистики»

*Баранов А.Н.*Объект и методы прикладной лингвистики // Баранов А.Н. Введение в прикладную лингвистику. М., 2003. С. 6-12

*Коваль С. А.* Лингвистические проблемы компьютерной морфологии. – СПб.: Изд-во С.-Петерб. ун-та, 2005. - 151 с. <http://skowal.narod.ru/TeachCompMorph.htm>. (Часть 1)

*Леонтьева Н.Н.* Автоматическое понимание текстов: системы, модели, ресурсы. М. 2006

*Маннинг К.Д., Рагхаван П., Шютце Х.* Глава 1. Булев поиск // Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск., М. 2011, стр. 23-40

*Маннинг К.Д., Рагхаван П., Шютце Х.* Глава 2. Лексикон и списки словопозиций // Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск., М. 2011, стр. 41-68

*Jurafsky D., Martin J.H.* (2000) Introduction // Jurafsky D., Martin J.H. Speech and language Processing, Prentice-Hall, , стр. 1-18

*Jurafsky D., Martin J.H.* (2000) Regular expressions and automata // Jurafsky D., Martin J.H. Speech and language Processing, Prentice-Hall, , стр. 22-31

*Jurafsky D., Martin J.H.* N-grams // Jurafsky D., Martin J.H. Speech and language Processing, Prentice-Hall, 2000, стр. 22-31

*Porter M.F*.An algorithm for suffix stripping, Program 14(3), 103-127

Раздел «Проектная работа над лингвистическими ресурсами»

*Paroubek P.* On the evaluation of the automatic parsing of natural language // Evaluation of text and speech systems. Text, speech and language technology. Vol. 37. Springer, 2007. P. 99–113.

*Segalovich I.* "A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine.", MLMTA-2003

Литература для проектной работы над лингвистическими ресурсами в том числе подбирается предподавателем индивидуально в рамках тематики проекта.

# Образовательные технологии

Студенты делятся на проектные группы. Выбор проекта делается совместно студентом и преподавателем. Основными критериями, определяющими выбор проекта являются а)релевантность темы для образовательной программы компьютерной лингвистики, б) актуальность темы, в том числе и для научной среды факультета в) востребованность результатов проекта научным сообществом, перспективы публикации результатов в) возможность выкладывания открытых данных по результатам проектов г) возможность привлечения внешних консультантов по проекту. Проектная группа формирует техническое задание, которое обсуждается и принимается всеми участниками семинара. Техническое задание служит индивидуальным планом работы каждой группы и задает основные рубежи отчетности. Отчетность производится в онлайновом режиме в форме представления текущих материалов и в оффлайн режиме в форме обсуждений данных и промежуточных результатов.

# Оценочные средства для текущего контроля и аттестации студента

## Тематика заданий текущего контроля

Домашнее задание 1: Подготовка технического задания проекта с указанием типов задач, ролей членов проектной команды и сроков обсуждения

Домашнее задание 2: Ведение еженедельного протокола выполнения поставленных задач на открытом проектном ресурсе в интернете.

Коллоквиумы 1-5:

Подготовка презентации хода работы над проектом: реферирование релевантной статьи по проекту, подготовка данных для интерактивного анализа, подготовка презентации промежуточных результатов, анализ сложностей и проблем.

## Вопросы для оценки качества освоения дисциплины

1. Расскажите, в чем состоит итоговая задача вашего проекта
2. Расскажите об основных выбранных методах решения
3. Расскажите об используемых данных
4. Расскажите о распределении ролей в проекте
5. Расскажите о проблемах, с которыми вы столкнулись при решении проектных задач,

## Примеры заданий промежуточного /итогового контроля

1. Изложите суть и задачу проекта
2. Напишите план реализации проекта
3. Подготовьте презентацию плана
4. Подготовьте презентацию данных, проблем, связанных с данными и их решение
5. Подготовьте исходную презентацию проекта
6. Подготовьте коллективный обзор на одну из тем предложенную преподавателем

# Учебно-методическое и информационное обеспечение дисциплины.

## Базовый учебник

*Jurafsky, Daniel, and James H. Martin. (2009).*[*Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*](http://www.cs.colorado.edu/~martin/slp.html)*. 2nd edition. Prentice-Hall*

## Основная литература

Маннинг К.Д., Рагхаван П., Шютце Х. Глава 1. Булев поиск // Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск., М. 2011

*Баранов А.Н.* Объект и методы прикладной лингвистики // Баранов А.Н. Введение в прикладную лингвистику. М., 2003. С. 6-12

*Коваль С. А.* Лингвистические проблемы компьютерной морфологии. – СПб.: Изд-во С.-Петерб. ун-та, 2005. - 151 с. <http://skowal.narod.ru/TeachCompMorph.htm>. (Часть 1)

*Леонтьева Н.Н.* Автоматическое понимание текстов: системы, модели, ресурсы. М. 2006

*Маннинг К.Д., Рагхаван П., Шютце Х.* Глава 1. Булев поиск // Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск., М. 2011, стр. 23-40

*Маннинг К.Д., Рагхаван П., Шютце Х.* Глава 2. Лексикон и списки словопозиций // Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск., М. 2011, стр. 41-68

*Jurafsky D., Martin J.H.* (2000) Introduction // Jurafsky D., Martin J.H. Speech and language Processing, Prentice-Hall, , стр. 1-18

*Jurafsky D., Martin J.H.* (2000) Regular expressions and automata // Jurafsky D., Martin J.H. Speech and language Processing, Prentice-Hall, , стр. 22-31

*Jurafsky D., Martin J.H.* N-grams // Jurafsky D., Martin J.H. Speech and language Processing, Prentice-Hall, 2000, стр. 22-31

*Porter M.F*.An algorithm for suffix stripping, Program 14(3), 103-127

## Дополнительная литература

*Баранов А.Н.* Оптимизация общения с ЭВМ: системы обработки естественного языка  // *Баранов А.Н*. Введение в прикладную лингвистику. М., 2003. С. 6-12

*Захаров* *В.П.* Информационно-поисковые системы: Учебно-метод. пособие.СПб., 2005.

Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2009 (Петрозаводск, 16 сентября 2009г.). Санкт-Петербург: НУ ЦСИ, 2009.

*Рубашкин В.Ш.* Прикладная лингвистика и языковая инженерия // Труды международной конференции «MegaLing’2005. Прикладная лингвистика в поиске новых путей». – СПб.: Изд-во «Осипов», 2005.

Структурная иприкладная лингвистика*. Под ред. А. С.* Герда*. Вып. 1. Л., 1978. — Вып. 7. СПб.,* 2008

*Фридл Дж.* Регулярные выражения. М. 2003

*Paroubek P.* On the evaluation of the automatic parsing of natural language // Evaluation of text and speech systems. Text, speech and language technology. Vol. 37. Springer, 2007. P. 99–113.

*Segalovich I.* "A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine.", MLMTA-2003

# Материально-техническое обеспечение дисциплины

Занятия должны проходить в компьютерных классах, необходим проектор.