**Правительство Российской Федерации**

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

Факультет

**Программа дисциплины** «Компьютерная лингвистика»

для направления

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](mailto:olesar@gmail.com)

Толдова С.Ю., к.ф.н, [toldova@yandex.ru](file:///C:\Users\lkoldina\Downloads\toldova@yandex.ru)

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 студентов направления подготовки/ специальности, обучающихся по магистерской программе «Компьютерная лингвистика» изучающих дисциплину Компьютерная лингвистика.

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 для подготовки магистра для магистерской программы «Компьютерная лингвистика» утвержденным в 2012г.

# Цели освоения дисциплины

Целями освоения дисциплины «Компьютерная лингвистика» являются знакомство с основными проблемами, современными методами и подходами в области компьютерной лингвистики, основными инструментами и технологиями, ключевыми задачами и диапазоном имеющихся решений. Курс закладывает теоретическую и практическую базу для разработки и тестирования различных модулей автоматического лингвистического анализа, автоматической обработки текста при разработке и тестировании различных модулей автоматической обработки текстов

# 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
* использования существующих сред для автоматического лингвистического анализа

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 к самостоятельному освоению новых методов исследования, изменению научного и научно-производственного профиля своей деятельности | СК-3 |  |  |
| Способен совершенствовать и развивать свой интеллектуальный и культурный уровень, строить траекторию профессионального развития и карьеры | СК-М 4 |  |  |
| Способен анализировать, верифицировать, оценивать полноту информации в ходе профессиональной деятельности, при необходимости восполнять и синтезировать недостающую информацию и работать в условиях неопределенности | СК-6 | Воспроизводит конкретные методы для решения задач компьютерной лингвистики, таких как автоматический морфологический и синтаксический анализ, извлечение информации; распознает стандартные проблемы при анализе языковых данных и подбирает соответствующую литературу, воспроизводит описанные в данной литературе алгоритмы | Чтение литературы по конкретным современным проблемам автоматической обработки текста и извлечения информации, составление реферата, обсуждение на семинарах |
| Cпособен вести профессиональную, в том числе научно-исследовательскую деятельность в международной среде | СК-8 | Воспроизводит методы автоматического анализа текста, использует научные разработки, представленные в статьях материалов современных конференций по компьютерной лингвистике, участвует в исследовательских группах, результаты работы которых принимают к рассмотрению на конференциях | Выполнение домашних заданий, чтение и обсуждение статей, представленных на международных конференциях, на семинарах, подготовка презентаций |
| Способен проводить анализ качества языковых данных, корпусов, систем, использующихся для автоматической обработки естественного языка | ПК-9 | Воспроизводит методы анализа качества корпусов и других языковых ресурсов; применяет методы анализа качества систем автоматического анализа текста на практике | Выполнение домашних заданий |
| Способен анализировать языковые данные с применением качественных и количественных методов, различных инструментальных средств | ПК-10 | Владеет качественными методами анализа данных, применяет методы функционального тестирования при качественном анализе данных, применяет основные количественные методы к анализу неструктурированных языковых данных больших объемов | Выполнение домашних заданий |
| Способен разрабатывать и совершенствовать электронные языковые ресурсы, лингвистические компоненты информационных систем, лингвистические модули компьютерных систем обучения | ПК-17 | Знает основные лингвистические ресурсы, владеет цепочкой лингвистической обработкой текста, умеет использовать среды разработки лингвистических систем | Выполнение домашних заданий, обсуждение на семинарах |
| Способен разрабатывать и совершенствовать электронные языковые ресурсы, лингвистические компоненты информационных систем, лингвистические модули компьютерных систем обучения | ПК-18 | Знает основные принципы разработки электронных ресурсов разного типа, разрабатывает собственные корпуса под конкретные задачи автоматической обработки текста; участвует в разработке основных лингвистических компонентов информационных систем | Выполнение домашнего задания по созданию и разметки собственных ресурсов в процессе выполнения домашних заданий |

# 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 теория 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 | Лингвистические ресурсы | 48 | 8 |  |  | 32 |
| 2 | Лингвистическая разметка | 48 | 12 |  |  | 32 |
| 3 | Инструменты для анализа и обработки текста | 48 | 4 |  |  | 32 |
| 4 | Основные направления современной компьютерной лингвистики | 144 | 20 | 20 | 24 | 104 |
|  | итого | 288 | 44 | 20 | 24 | 200 |

# Формы контроля знаний студентов

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | | | 2 год | | Параметры \*\* |
| 1 | 2 | 3 | 1 | 2 |  |
| Текущий  (неделя) | Домашнее задание |  | 2 нед. |  |  |  | Реферирование статей. Подготовка корпуса. Проверка на семинарских занятиях |
| Домашнее задание |  |  |  | 2 |  | 2 |
| Коллоквиум по проектам |  | 1 |  |  |  | Проверка на занятии |
| Контрольная работа |  |  |  |  | 1 | Проверка на занятии |
|  |  |  |  |  |  |  |  |
| Промежуточный | Экзамен |  |  | 3 |  |  | письменный экзамен в третьем модуле: длительность 120 мин. |
| Итоговый |  |  |  |  |  | 2 | письменный экзамен во втором модуле: длительность 120 мин. |

## Критерии оценки знаний, навыков

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

Домашнее задание по подготовке корпуса является групповым проектом с индивидуальными заданиями, выполняемыми в рамках проекта. Подготовку корпуса оценивается как степень соответствия выполненных индивидуальных заданий по подготовке корпуса и по его разметке стандартам и требованиям сбора текстов, метаразметки и других типов аннотирования корпусов (соответствие требований к разработке электронных ресурсов).

## 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 правильность 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*.

Накопленная оценка за текущий контроль учитывает результаты студента по текущему контролю следующим образом:

*Отекущий* = *0,6·Одз + 0,4·Окол*;

Способ округления накопленной оценки текущего контроля: в пользу студента.

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

*Опромежуточный = k1·Оэкзамен + k2· Отекущий + 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 = 0,5*

*k2 = 0,3*

*k3 = 0,2*

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

В диплом выставляет результирующая оценка по учебной дисциплине, которая формируется по следующей формуле:

*Одисциплина = 0,5·Опромежуточный* + 0,5·*Оитоговый*

Способ округления результирующей оценки по учебной дисциплине: в пользу студента.

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

# Содержание дисциплины

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Название | лекции | практические занятия | Литература или сетевые ресурсы по разделу |
| Раздел 1. Лингвистические ресурсы | | | | |
| 1 | Лингвистические данные | 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  О. Н. Ляшевская, С. А. Шаров, Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.  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.  [Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце](http://www.ozon.ru/context/detail/id/5497130/#tab_person) Введение в информационный поиск, М. 2011, Вильямс  <http://wordnet.princeton.edu/>  <https://framenet.icsi.berkeley.edu/fndrupal/> |
| 2 | Корпуса как вид линвистических данных |  | 2 |
| 3 | Электронные словари, словарное представление корпусных данных | 2 |  |
| 4. | Технологии “bag of words” в документе и корпусе | 2 |  |
| 5. | Ресурс Wordnet и его возможности | 2 |  |
| 6. | Ресурс Фреймнет и его возможности | 2 | 2 |
| 7. | Практическое занятие по собранным корпусам |  | 2 |
| Раздел 2. Лингвистическая разметка | | | | |
| 1. | Стеммер Портера |  | 2 | [Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце](http://www.ozon.ru/context/detail/id/5497130/#tab_person) Введение в информационный поиск, М. 2011, Вильямс,  Коваль С. А. Лингвистические проблемы компьютерной морфологии. - СПб.: Изд-во С.-Петерб. ун-та, 2005. - 151 с. <http://skowal.narod.ru/TeachCompMorph.htm>. (Часть 2)  Сокирко А. В. Морфологические модули на сайте www.aot.ru //Материалы конференции "Диалог-2004". Эл. версия: <http://www.dialog-21.ru/Archive/2004/Sokirko.htm>  Материал Ромип с сайта <http://romip.ru/> |
| 2. | Морфологические парсеры | 4 |  |
| 3. | Механизмы дизамбигуации | 4 |  |
| 4. | Оценка качества морфологической разметки |  | 2 |
| 5 | Технологии shallow parsing | 2 |  |
| 6. | Технологии chunking | 2 |  |
| 7. | Синтаксические парсеры |  | 4 |
| 8 | Оценка качества синтаксической разметки |  | 4 |
| Раздел 3.  Инструменты для анализа и обработки текста | | | | |
| 1. | среда разработчика лингвистических компонентов NLTK | 2 | 2 | Ресурсы с прилагающейся на сайтах документацией:  <http://gate.ac.uk/> <http://www.nooj4nlp.net/pages/nooj.html>  <http://uima.apache.org/>  <http://nltk.org/> |
| 2 | среда разработчика лингвистических компонентов GATE | 2 | 2 |
| Раздел 4.  Основные направления современной компьютерной лингвистики | | | | |
| 1 | Статистические методы в семантике | 2 | 2 | 1) Martin & Jurafsky "Speech & Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. 2009. Ch. 15-16  2) Manning C., Schutze H. Collocations // Manning C., Schutze H.Foundations of Statictical Natural Language Processing, 2002  Jurgens D., - An Evaluation of Graded Sense Disambiguation usingWord Sense Induction // First Joint Conference on Lexical and Computational Semantics, pp. 189–198. Navigli R. - Word Sense Disambiguation: A Survey //ACM Computing Surveys, 41(2), 2009, pp. 1–69 <http://promethee.philo.ulg.ac.be/engdep1/download/bacIII/ACM_Survey_2009_Navigli.pdf>  Ponzetto S., Navigli R. -   Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics , 2010, pp 1522-1531  Thanh Phong Pham, Hwee Tou Ng, Wee Sun Lee – Word Sense Disambiguation with Semi-Supervised Learning // American Association for Artificial Intelligence, 2005  <http://www.comp.nus.edu.sg/~nght/pubs/aaai05_wsd_ssup.pdf> |
| 2 | Коллокации.. Автоматическое распознавание семантических ролей (semantic role labeling) | 2 | 2 | Manning C., Schutze H. Collocations // Manning C., Schutze H.Foundations of Statictical Natural Language Processing, 2002, pp.151-189 |
| 3 | Анафора и кореференция   |  | | --- | |  | | 2 | 2 | Martin & Jurafsky "Speech & Language Processing. Chapter 21: Computational Discourse. Разделы: 21.3-21.8 Mitkov, R. (1999): *Anaphora resolution: the state of the art*, Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.6235&rep=rep1&type=pdf>  Iida, Ryu and Inui, Kentaro and Matsumoto, Yuji. Anaphora Resolution by Antecedent Identification Followed by Anaphoricity Determination, 2005. <https://dl.acm.org/citation.cfm?id=1113308.1113312>  П.В. Толпегин и другие. Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения, 2006. <http://www.dialog-21.ru/digests/dialog2006/materials/html/Tolpegin.htm>  Деликишкина Е.А., Федорова О.В. Влияние фактора синтаксической роли антецедента на разрешение референциальной неоднозначности в русском языке, 2012. <http://www.dialog-21.ru/digests/dialog2012/materials/pdf/92.pdf> |
| 4 | Анализ дискурса | 2 | 2 | Mann, W.C., & Thompson, S.A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization <http://www.cis.upenn.edu/~nenkova/Courses/cis700-2/rst.pdf>  D.Marcu & A.Echihabi An Unsupervised Approach to Recognizing Discourse Relations <http://acl.ldc.upenn.edu/P/P02/P02-1047.pdf> |
| 5 | Извлечение фактов и отношений | 2 | 2 | Martin & Jurafsky "Speech & Language Processing. Chapter 22. Разделы 22.1 (Named Entity Recognition) и 22.2 (Relation Detection and Classification). Только в издании 2008 года!  Nadeau, Sekine, - A survey of named entity recognition and classification// Linguisticae Investigationes 30(1):3--26 , 2007  <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>  Sun, Grishman, Sekine, - Semi-supervised Relation Extraction with Large-scale Word Clustering// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics , pages 521–529,Portland, Oregon, June 19-24, 2011.  [http://aclweb.org/anthology//P/P11/P11-1053.pdf](http://aclweb.org/anthology/P/P11/P11-1053.pdf)  Chan, Roth, - Exploiting Syntactico-Semantic Structures for Relation Extraction// HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. Pages 551-560 |
| 6 | Тональность и извлечение мнений | 2 | 2 | <http://www.dialog-21.ru/digests/dialog2011/materials/en/pdf/50.pdf>  <http://www.dialog-21.ru/digests/dialog2013/materials/pdf/PanichevaPV.pdf> и <http://www.dialog-21.ru/digests/dialog2013/materials/pdf/FrolovAV.pdf>  <http://www.dialog-21.ru/digest/2013/pdf1/>  "Использование метода условных случайных полей для обработки текстов на русском языке"  <http://nlp.stanford.edu/sentiment/index.html>  <http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf> |
| 7 | Автоматическая генерация текста | 2 | 2 | Обзор:<http://www.dialog-21.ru/Archive/2004/Sokolova.htm>  Статистическая генерация:  <http://cs.stanford.edu/~pliang/papers/generation-emnlp2010.pdf>  [http://aclweb.org/anthology//P/P13/P13-1138.pdf](http://aclweb.org/anthology/P/P13/P13-1138.pdf)  [http://aclweb.org/anthology//C/C10/C10-1012.pdf](http://aclweb.org/anthology/C/C10/C10-1012.pdf) |
| 8 | Диалоговые системы | 4 | 4 | <https://en.wikipedia.org/wiki/Readability>  <http://www.dsusd.us/users/christopherg/measuring%20reading%20comprehension.pdf>  <http://dl.acm.org/citation.cfm?id=2063639> |

# Образовательные технологии

Рекомендуемые образовательныетехнологии включают лекции, коллоквиумы, самостоятельную работу студентов (чтение и реферирование литературных источников, в том числе англоязычных, выполнение практических домашних заданий с использованием специализированного компьютерного инструментария, курсовые работы (по выбору обучающихся), предусмотрены также различные виды производственных практик, обладающих необходимым кадровым и научно-техническим потенциалом, а также в сторонних организациях: профильных вузах и предприятиях, институтах РАН, НИИ, компаниях.

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.

В рамках учебного курса могут быть предусмотрены встречи с российскими или зарубежными учеными, специализирующимися в области создания корпусов.

# Оценочные средства для текущего контроля и аттестации студента

## Тематика заданий текущего контроля

Примерные вопросы/ задания для домашних работ:

1. Создайте список наиболее частотных терминов вашего корпуса
2. Постройте ветку типа wordnet для заданных слов с учетом данных НКРЯ и собранного вами корпуса
3. Проведите морфологическую разметку корпуса
4. Оцените качество предложенных результатов морфологической/синтаксической разметки

## 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. Каковы методы взвешивания релевантности слов и словосочетаний в корпусе?
5. Ресурс типа wordnet, его свойства
6. Ресурс типа framnet, его свойства
7. Принципы работы морфологических парсеров
8. Принципы работы стеммера Портера
9. Методы снятия неоднозначности
10. Основания оценки качества автоматического морфологического разбора
11. Технология shallow parcing
12. Технология chunking
13. Принципы работы синтаксических парсеров
14. Основания оценки качества автоматического синтаксического разбора
15. Особенности и возможности среды GATE
16. Особенности и возможности среды NLTK
17. Современные подходы к автоматическому разрешению неоднозначности
18. Колокации. Автоматическая разметка семантических ролей
19. Автоматический анализ анафоры и кореференции
20. Автоматический анализ дискурса
21. Тональность и извлечение мнений
22. Извлечение фактов и отношений
23. Автоматическая генерация текста
24. Диалоговые системы

# 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

[Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце](http://www.ozon.ru/context/detail/id/5497130/#tab_person) Введение в информационный поиск, М. Вильямс, 2011,

## Дополнительная литература

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

Искусственный интеллект: Справочник: Кн.1: Системы общения и экспертные системы. - М.: Радио и связь, 1990.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. Лингвистический процессор для сложных информационных систем

## Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства:

* Программные среды GATE, NLTK, NooJ, UIMA

## Дистанционная поддержка дисциплины

Для освоения программы используются электронные ресурсы Wordnet, Framenet

# Материально-техническое обеспечение дисциплины

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