**Правительство Российской Федерации**

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

Факультет филологии

**Программа дисциплины** Компьютерная лингвистика

для направления 035800.62 Фундаментальная и прикладная лингвистика   
подготовки бакалавра

Авторы программы:

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

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

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, 2016

*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 для направления 035800.62 Фундаментальная и прикладная лингвистика подготовки бакалавра изучающих дисциплину Компьютерная лингвистика.

Программа разработана в соответствии с:

* Стандартом НИУ ВШЭ по направлению подготовки 035800.62 «Фундаментальная и прикладная лингвистика» подготовки бакалавров;
* Рабочим учебным планом университета по направлению подготовки 035800.62 «Фундаментальная и прикладная лингвистика» подготовки бакалавров утвержденным в 2012г.

# Цели освоения дисциплины

Целями освоения дисциплины «Компьютерная лингвистика» являются знакомство с основными проблемами в области компьютерной лингвистики, базовыми алгоритмами, основными инструментами и технологиями в области автоматической обработки естественного языка. 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 задачи компьютерной лингвистики;
* основные формальные модели, лежащие 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;
* принципы оценки качества таких систем.

*Уметь*

* создавать модули первичной обработки текста;
* строить формальную модель морфологии для создания системы автоматического морфологического анализа;
* проводить оценку качества систем автоматического морфологического, синтаксического и семантического анализа;
* использовать соответствующие модули в различных приложениях.

*Иметь навыки (приобрести опыт)*

* разработки программ первичной обработки текста
* использования систем автоматического морфологического анализа
* тестирования систем морфологического и синтаксического 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 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 | - чтение специальной литературы  -выполнение самостоятельных заданий  - анализ полученных данных |
| 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 | практические занятия по созданию языковых ресурсов и лингвистических компонентов  обзор и рефераты существующих разработок |

# 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 алгоритмов и основ математики
* 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. Основные задачи компьютерной лингвистики | 12 | 2 | 2 |  | 8 |
| 2 | Первичная обработка текста. Графематический анализ | 24 | 4 | 4 |  | 16 |
| 3 | Автоматический морфологический анализ | 36 | 6 | 6 |  | 24 |
| 4 | Автоматический синтаксический анализ | 36 | 6 | 6 |  | 24 |
|  |  | 108 | 18 | 18 |  | 72 |

# Формы контроля знаний студентов

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля |  | | | | Параметры \*\* |
|  | 2 | 3 |  |
| Текущий  (неделя) | Домашнее задание |  | 2 | 4 |  | Реферирование статей. Проверка на семинарских занятиях |
| Домашнее задание |  | 2 | 4 |  | Выполнение практических заданий. Проверка на семинарских занятиях |
|  | Проектная работа |  | 2 | 1 |  | Коллоквиум по проектам. Проверка на занятии |
|  | Подготовка презентации |  | 1 |  |  | Презентация по теме. Проверка на занятии |
| Итоговый | Экзамен |  |  |  |  | Устный экзамен, защита проекта |

## Критерии оценки знаний, навыков

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

Домашнее задание по подготовке корпуса, разработке программы предварительного анализа текста, тестированию морфологических анализаторов являются групповыми проектами с индивидуальными заданиями, выполняемыми в рамках проектов.

Тестирование программы предварительного анализа текста и морфологических анализаторов проходит в формате Форума по оценке систем автоматической обработке текста. Командам выдается тестовый и эталонный корпус. Каждая команда проводит оценку точности и полноты, а также функциональное тестирование соответствующей программы.

Задания и проверка происходит через LMS. Задания выполняются в формате тестов и проектов в LMS.

## Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на семинарских и практических занятиях: оцениваются вопросы к прочитанному материалу, качество презентации, активность при проведении тестирования систем. 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 - *Оаудиторная*.

Преподаватель оценивает самостоятельную работу студентов: оценивается правильность и своевременность сдачи домашних работ, качество задаваемых к прочитанному материалу вопросов и составление резюме по прочитанным статьям. 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:

О*накопленная*= 0.4\* *Отекущий* + 0.2\* Оауд + 0.4\* Осам.работа

где*Отекущий* рассчитывается как взвешенная сумма всех форм текущего контроля, предусмотренных в РУП

*Отекущий* = *0.8·Окол + 0.2·Одз* ;

Способ округления накопленной оценки текущего контроля: способ округления - в пользу студента.

Результирующая оценка за дисциплину рассчитывается следующим образом:

*Орезульт = 0.7\* Онакопл + 0.3 \*·Оэкз/зач*

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

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

На зачете студент может получить дополнительный вопрос (дополнительную практическую задачу, решить к пересдаче домашнее задание), ответ на который оценивается в 1 балл.

# Содержание дисциплины

|  |  |  |
| --- | --- | --- |
|  | лекции | практические занятия |
| Раздел1. Введение.  Основные задачи компьютерной лингвистики |  |  |
| 1. Введение в компьютерную лингвистику  2. Задачи компьютерной лингвистики  3. **Коллоквиум:** лингвистические системы | 2 | 2 |
| Раздел 2. Первичная обработка текста.  Графематический анализ |  |  |
| 4. Поиск, индексация, оценка качества  5.-6. Задачи, этапы и технологии первичной обработки текста. Проблемы токенизации.  7. Анализ различных подходов к токенизации и делению на предложение.  8.-9. Оценка качества токенизаторов (коллоквиум, отчет по проекту, оценка качества) | 2  2 | 2  2 |
| Раздел 3. Автоматический морфологический анализ |  |  |
| 10. Автоматический морфологический анализ. Введение  11. Морфологический анализ: организация данных  12. Конечные автоматы и конечные преобразователи в морфологическом анализе | 2  2 | 2 |
| 13.-14. Методы дизамбигуации (машинное обучение: rule-induction, HMM). | 2 | 2 |
| 15. **Коллоквиум.** Оценка качества морфологических анализаторов |  | 2 |
| Раздел 4. Автоматический синтаксический анализ |  |  |
| 16. Формализмы и методы автоматического синтаксического анализа | 2 | 2 |
| 17. Алгоритмы автоматического синтаксического анализа. Анализ работы систем | 2 | 2 |
| 18. Синтаксический анализ в терминах деревьев зависимостей | 2 | 2 |
| Экзамен |  |  |

### Литература

|  |
| --- |
| Раздел1. Введение. Основные задачи компьютерной лингвистики |
| 1. Введение в компьютерную лингвистику  Прикладная лингвистика. // Статья в энциклопедии «Фонд знаний «Ломоносов»». URL: <http://www.lomonosov-fund.ru/enc/ru/encyclopedia:01206:article>.  Литература для докладов (презентации на коллоквиуме):  *Генерация текстов*  Соколова Е.Г., Болдасов М. Автоматическая генерация текстов на ЕЯ (портрет направления)<http://www.dialog-21.ru/Archive/2004/Sokolova.htm>  Ehud Reiter. Has a Consensus NL Generation Structure Appeared, and is it Psycholinguistically Plausible? — 7th International Generation Workshop (Kennebunkport, Maine).  URL: <http://www.aclweb.org/anthology/W/W94/W94-0319.pdf> *Машинный перевод* Лекция Л.Л.Иомдина “Машинный перевод: успехи, неудачи, надежды”. Лекторий Политехнического музея. Видео. Доступно с URL <http://rutube.ru/video/828268c50a49b876a6f4676b839fa745/> дата обращения 20.01.2014)  3. Brown P. F. et al. The mathematics of statistical machine translation: Parameter estimation //Computational linguistics. – 1993. – Т. 19. – №. 2. – С. 263-311. <http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf> *Text mining (классификация, кластеризация, реферирование). На примере анализа новостного потока* Кондратьев М. Е. Анализ методов кластеризации новостного потока //Тр. Восьмой Всерос. науч. конф.«Электронные библиотеки: перспективные методы и технологии, электронные коллекции»(RCDL’2006).—Ярославль. – 2006. – С. 108-114.rcdl.ru/doc/2006/paper\_92\_v1.pdf‎ *Распознавание речи* Speech recognition. http://en.wikipedia.org/wiki/Speech\_recognition  В. Н. Сорокин, *Синтез речи*. М., 1992.,  D. Childers, *Speech Processing*, John Wiley and Sons, Inc., все издания, начиная с 1999 *Диалоговые системы. Говорящие роботы* Filipe P. P., Morgado L., Mamede N. J. An Adaptive Domain Knowledge Manager for Dialogue Systems //ICEIS (5). – 2007. – С. 45-52. <http://www.inesc-id.pt/pt/indicadores/Ficheiros/3983.pdf>  Bermúdez M. G., Vila M. G. Dialogue Management for multilingual communication through different channels. *Извлечение информации из текста: извлечение именованных сущностей, отношений и фактов* Information Extraction  <http://en.wikipedia.org/wiki/Information_extraction>  Котельников Д. С., Лукашевич Н. В. Итерационное извлечение шаблонов описания событий по новостным кластерам. <http://ceur-ws.org/Vol-934/paper45.pdf>  Nadeau D. and Sekine S. A survey of named entity recognition and classification, Linguisticae Investigationes, Amsterdam, Netherlands: John Benjamins Publishing Company, 1: Vol. 30. pp. 3‑26.  Brykina M., Toldova S.Yu., Faynveyts A. V. Dictionary-based ambiguity resolution in Russian named-entities recognition. A case study. P. 163-177. Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции “Диалог” 2013. Вып. 12(19). М.: РГГУ, 2013. URL: <http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BrykinaMM.pdf>  Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке //ЗАО «Ай-Теко», Москва, Россия. URL: <http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/50.pdf> |
| Раздел 2. Первичная обработка текста.  Графематический анализ |
| 4. Поиск, индексация, оценка качества  [Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце](http://www.ozon.ru/context/detail/id/5497130/#tab_person). Введение в информационный поиск, М. 2011, Вильямс Гл. 1 пп.1.1.-1.2., гл.2. пп. 2.1-2.2.  5.-6. Задачи, этапы и технологии первичной обработки текста. Проблемы токенизации.  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.  Автоматическая обработка текста. Графематика. http://www.aot.ru/docs/graphan.html  7. Анализ различных подходов к токенизации и делению на предложение.  O.Урюпина. Автоматическое разбиение текста на предложения для русского языка. Результат: три вопроса к тексту <http://www.dialog-21.ru/digests/dialog2008/materials/html/83.htm> |
| Раздел 3. Автоматический морфологический анализ |
| 10. Автоматический морфологический анализ. Введение  Коваль С. А. Лингвистические проблемы компьютерной морфологии. - СПб.: Изд-во С.-Петерб. ун-та, 2005. - 151 с.  11. Морфологический анализ: организация данных  12. Конечные автоматы и конечные преобразователи в морфологическом анализе  Jurafsky D., James H. Regular expressions and automata (2.2 Finite State automata). // Jurafsky D., James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. – 2009. <http://www.cs.vassar.edu/~cs395/docs/2.pdf>  Jurafsky D., James H. Words and transducers. // Jurafsky D., James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. – 2009. <http://people.mokk.bme.hu/~kornai/termeszetes/3.pdf> |
| 13.-14. Методы дизамбигуации (машинное обучение: rule-induction, HMM).  Jurafsky D., James H. N-grams. Ch.8. 8.1-8.5 // Jurafsky D., James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. Ch. 8. N-Grams. 2009.  <http://web.mit.edu/6.863/www/fall2012/readings/ngrampages.pdf>  Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). URL: <http://www.aot.ru/docs/RusCorporaHMM.htm> ,  Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. // Компьютерная лингвистика и интеллектуальные технологии. // Труды международного семинара Диалог’2005. – М., 2005.  URL: <http://www.dialog-21.ru/Archive/2005/Zelenkov%20Segalovich/Zelenkov_Segalovich.htm>  (http://download.yandex.ru/company/Zelenkov\_Segalovich.pdf),  Екатерина Протопопова, Виктор Бочаров. Автоматическое извлечение правил для снятия морфологической неоднозначности. // В кн.: Доклады всероссийской научной конференции АИСТ’2013 / Отв. ред.: Е. Л. Черняк; науч. ред.:Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ», 2013. С. 184-95.  URL: http://aistconf.org/stuff/aist2013/submissions/aist2013\_submission\_28.pdf  Лакомкин Е. Д., Рыжова Д. А., Пузыревский И. Анализ статистических алгоритмов снятия морфологической омонимии в русском языке // В кн.: Доклады всероссийской научной конференции АИСТ’2013 / Отв. ред.: Е. Л. Черняк; науч. ред.:Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ», 2013. С. 184-195.  URL: http://aistconf.org/stuff/aist2013/submissions/aist2013\_submission\_33.pdf |
| 15. Коллоквиум. Оценка качества морфологических анализаторов  *Системы для оценки:*   |  |  | | --- | --- | | * Mystem (компании Яндекс) | [http://company.yandex.ru/technologies/mystem/](https://www.google.com/url?q=http://company.yandex.ru/technologies/mystem/&usd=2&usg=ALhdy2_oVM_1MKNHhmYvHeYv66ksHDAObQ) | | * Pymorphy | [http://pythonhosted.org/pymorphy/](https://www.google.com/url?q=http://pythonhosted.org/pymorphy/&usd=2&usg=ALhdy28KkHIYC8OclkVvBxM-xXRkm3vIIQ) | | * АОТ | [http://www.aot.ru/download.php](https://www.google.com/url?q=http://www.aot.ru/download.php&usd=2&usg=ALhdy29fwXE0B1gDQ6Z4ctacYO0p6pSvLQ) | | * FreeLing | [http://nlp.lsi.upc.edu/freeling/](https://www.google.com/url?q=http://nlp.lsi.upc.edu/freeling/&usd=2&usg=ALhdy28kVYbLKOTh6fSPkuI6UujSqUt8LA) | | * TreeTagger | [http://corpus.leeds.ac.uk/mocky/](https://www.google.com/url?q=http://corpus.leeds.ac.uk/mocky/&usd=2&usg=ALhdy28Au47axc71fwZo96R-OPl29IgtRw) | | * TnT | [http://corpus.leeds.ac.uk/mocky/](https://www.google.com/url?q=http://corpus.leeds.ac.uk/mocky/&usd=2&usg=ALhdy28Au47axc71fwZo96R-OPl29IgtRw) |   *Дополнительное чтение по разделу:*  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> |
| Раздел 4. Автоматический синтаксический анализ |
| 16. Формализмы и методы автоматического синтаксического анализа  Jurafsky D., James H. Chapter 13. Parsing with Context-free Grammar. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. – 2009. The 2nd edition (Chapter 9. Edition - 2000). |
| 17. Алгоритмы автоматического синтаксического анализа. Анализ работы систем |
| 18. Синтаксический анализ в терминах деревьев зависимостей  Апресян Ю. Д. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы //Национальный корпус русского языка. – 2003. – Т. 2005. – С. 193-214. URL: http://corpora.phil.spbu.ru/Works2008/Boguslavsky1\_56\_74.pdf (дата обращения: 27.01.2015)  Анисимович К. В. и др. Синтаксический и семантический парсер, основанный на лингвистических технологиях ABBYY Compreno. // В кн. Компьютерная лингвистика и интеллектуальные технологии. (дата обращения: 27.01.2015)  По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций – (URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/anisimovich.pdf (дата обращения: 27.01.2015))  Иомдин Л.Л. и др. Синтаксический анализатор системы ЭТАП: современное состояние. // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций  – (URL: dialog2012/materials/pdf/Iomdin.pdf (дата обращения: 27.01.2015))  Антонова А. А., Мисюрев А. В. Анализатор русского языка syntautom для соревнования синтаксических парсеров (диАлог-2012). // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций. (http://www.dialog-21.ru/digests/dialog2012/materials/pdf/98.pdf (дата обращения: 27.01.2015))  *Корпуса для анализа:*   * + СинТагРус http://ruscorpora.ru/search-syntax.html (дата обращения: 27.01.2015),   + Тестовый корпус с параллельной синтаксической разметкой http://otipl.philol.msu.ru/~soiza/testsynt/, (дата обращения: 27.01.2015),   + Rus-Treebank http://otipl.philol.msu.ru/~soiza/rtb/res01/rtb.php (дата обращения: 27.01.2015) |

# 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. Разбейте текст на токены и предложения. Составьте список необходимых для сегментации классов символов, типов токенов; составьте список сложных случаев, предложите решение.
3. Задайте три вопроса к прочитанной статье (главы из учебника, посвященные формализмам, используемым в автоматической обработке текста; статьи, посвященные реальным системам и методам автоматического анализа текстов)
4. Протестируйте систему сегментации текста
5. Постройте конечный автомат и конечный преобразователь для описания правил морфонологических чередований и построения словоформ на одном из предложенных языков
6. Проведите морфологическую дизамбигуацию некоторого текста. Какие типы омонимии вам встретились. Оцените качество предсказаний системы.
7. Постройте контекстно-свободную грамматику для анализа некоторой синтаксической конструкции
8. Предложите синтаксическую разметку предложения, основанную на деревьях зависимости
9. Оцените качество предложенных результатов морфологической/синтаксической разметки

## Вопросы для оценки качества освоения дисциплины

1. Какие типы лингвистических данных вам известны?
2. Принципы работы морфологических парсеров
3. Конечные автоматы и конечные преобразователи в морфологическом анализе
4. Методы снятия морфологической неоднозначности
5. Основания оценки качества автоматического морфологического разбора
6. Технология shallow parcing
7. Технология chunking
8. Принципы работы синтаксических парсеров
9. Основания оценки качества автоматического синтаксического разбора

## Примеры заданий промежуточного /итогового контроля

По желанию автора программы, приводятся примеры билетов с вопросами и задачами, заданий для зачета или экзамена, тренировочные тесты по дисциплине.

1. Ответить на один из теоретических вопросов (п. 9.2.)
2. Построить конечный преобразователь для задачи морфологического анализа/синтеза словоформы класса Х в языке Y
3. Снять морфологическую неоднозначность. Ответ прокомментировать
4. Построить дерево зависимости для предложения Х. Оценить разбор данного предложения системой Y
5. Построить контекстно-свободную грамматику для конструкции Х.

# 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.

# Учебно-методическое и информационное обеспечение дисциплины

## Основная литература

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. Вильямс, 2011,

## Дополнительная литература

1. Автоматическая обработка текста. Графематика. http://www.aot.ru/docs/graphan.html
2. Анисимович К. В. и др. Синтаксический и семантический парсер, основанный на лингвистических технологиях ABBYY Compreno. // В кн. Компьютерная лингвистика и интеллектуальные технологии. (дата обращения: 27.01.2015)
3. Антонова А. А., Мисюрев А. В. Анализатор русского языка syntautom для соревнования синтаксических парсеров (диАлог-2012). // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций. (http://www.dialog-21.ru/digests/dialog2012/materials/pdf/98.pdf (дата обращения: 27.01.2015))
4. Апресян Ю. Д. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы //Национальный корпус русского языка. – 2003. – Т. 2005. – С. 193-214. URL: http://corpora.phil.spbu.ru/Works2008/Boguslavsky1\_56\_74.pdf (дата обращения: 27.01.2015)
5. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. Лингвистический процессор для сложных информационных систем
6. Екатерина Протопопова, Виктор Бочаров. Автоматическое извлечение правил для снятия морфологической неоднозначности. // В кн.: Доклады всероссийской научной конференции АИСТ’2013 / Отв. ред.: Е. Л. Черняк; науч. ред.:Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ», 2013. С. 184-95. URL: http://aistconf.org/stuff/aist2013/submissions/aist2013\_submission\_28.pdf
7. Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог’2005. – М., 2005. URL: <http://www.dialog-21.ru/Archive/2005/Zelenkov%20Segalovich/Zelenkov_Segalovich.htm> (http://download.yandex.ru/company/Zelenkov\_Segalovich.pdf),
8. Иомдин Л.Л. и др. Синтаксический анализатор системы ЭТАП: современное состояние. // В кн. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций – (URL: dialog2012/materials/pdf/Iomdin.pdf (дата обращения: 27.01.2015))
9. Искусственный интеллект: Справочник: Кн.1: Системы общения и экспертные системы. - М.: Радио и связь, 1990.
10. Коваль С. А. Лингвистические проблемы компьютерной морфологии. - СПб.: Изд-во С.-Петерб. ун-та, 2005. - 151 с.
11. [Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце](http://www.ozon.ru/context/detail/id/5497130/#tab_person). Введение в информационный поиск, М. 2011, Вильямс Гл. 1 пп.1.1.-1.2., гл.2. пп. 2.1-2.2.
12. Лакомкин Е. Д., Рыжова Д. А., Пузыревский И. Анализ статистических алгоритмов снятия морфологической омонимии в русском языке // В кн.: Доклады всероссийской научной конференции АИСТ’2013 / Отв. ред.: Е. Л. Черняк; науч. ред.:Д. И. Игнатов, М. Ю. Хачай, О. Баринова. М. : Национальный открытый университет «ИНТУИТ», 2013. С. 184-195. URL: <http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf>
13. По материалам ежегодной Международной конференции "Диалог" (2012). Том 2. Доклады специальных секций – (URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/anisimovich.pdf (дата обращения: 27.01.2015))
14. Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). URL: http://www.aot.ru/docs/RusCorporaHMM.htm,
15. Структурная и прикладная лингвистика. Под ред. А. С. Герда. Вып. 1. Л., 1978. — Вып. 7. СПб., 2008
16. Урюпина. O. Автоматическое разбиение текста на предложения для русского языка. Результат: три вопроса к тексту <http://www.dialog-21.ru/digests/dialog2008/materials/html/83.htm>

## Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства:

|  |
| --- |
| [http://company.yandex.ru/technologies/mystem/](https://www.google.com/url?q=http://company.yandex.ru/technologies/mystem/&usd=2&usg=ALhdy2_oVM_1MKNHhmYvHeYv66ksHDAObQ) |
| [http://pythonhosted.org/pymorphy/](https://www.google.com/url?q=http://pythonhosted.org/pymorphy/&usd=2&usg=ALhdy28KkHIYC8OclkVvBxM-xXRkm3vIIQ) |
| [http://www.aot.ru/download.php](https://www.google.com/url?q=http://www.aot.ru/download.php&usd=2&usg=ALhdy29fwXE0B1gDQ6Z4ctacYO0p6pSvLQ) |
| [http://nlp.lsi.upc.edu/freeling/](https://www.google.com/url?q=http://nlp.lsi.upc.edu/freeling/&usd=2&usg=ALhdy28kVYbLKOTh6fSPkuI6UujSqUt8LA) |
| [http://corpus.leeds.ac.uk/mocky/](https://www.google.com/url?q=http://corpus.leeds.ac.uk/mocky/&usd=2&usg=ALhdy28Au47axc71fwZo96R-OPl29IgtRw) |
| [http://corpus.leeds.ac.uk/mocky/](https://www.google.com/url?q=http://corpus.leeds.ac.uk/mocky/&usd=2&usg=ALhdy28Au47axc71fwZo96R-OPl29IgtRw) |

Один из пакетов на выбор (для группы)

* Программные среды GATE, NLTK, NooJ, UIMA

## Дистанционная поддержка дисциплины

Для освоения программы используются электронные ресурсы LMS,

|  |
| --- |
| [http://text0.mib.man.ac.uk:8080/scottpiao/sent\_detector](https://www.google.com/url?q=http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector&usd=2&usg=ALhdy2_oRZA0Gvbx80Jle7tWyOHGEnm99g) |
| [http://beta.visl.sdu.dk/visl/en/parsing/automatic/parse.php](https://www.google.com/url?q=http://beta.visl.sdu.dk/visl/en/parsing/automatic/parse.php&usd=2&usg=ALhdy2-XH81TnL0jEalsJTUGcQGKeRoMPw) |
| [http://ruscorpora.ru/](https://www.google.com/url?q=http://ruscorpora.ru/&usd=2&usg=ALhdy29fWCWUjFubgNuuWmngm92GuhXPFg) |
| [http://www.corpus-i.compling.net/res01/rtb.php](https://www.google.com/url?q=http://www.corpus-i.compling.net/res01/rtb.php&usd=2&usg=ALhdy283HSO1Y3FbuDBuZsPnxhp-IGhLnw) |
| [http://www.connexor.fi/](https://www.google.com/url?q=http://www.connexor.fi/&usd=2&usg=ALhdy28Xl16opZ_ttkIFSw7acZsY2eyStw) |
|  |

* СинТагРус http://ruscorpora.ru/search-syntax.html (дата обращения: 27.01.2015),
* Тестовый корпус с параллельной синтаксической разметкой http://otipl.philol.msu.ru/~soiza/testsynt/, (дата обращения: 27.01.2015),
* Rus-Treebank http://otipl.philol.msu.ru/~soiza/rtb/res01/rtb.php (дата обращения: 27.01.2015)

# Материально-техническое обеспечение дисциплины

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