МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Методические указания

по выполнению практических работ студентов по дисциплине «Корпусная лингвистика»

Направление подготовки	45.04.02 Лингвистика
Направленность (профиль)	Современные методы прикладной лингвистики и перевода
Год начала обучения	2026
Форма обучения	очная
Реализуется в семестре	1

Ставрополь
2026

Введение

Методические рекомендации по выполнению практических работ по дисциплине «Корпусная лингвистика» разработаны в соответствии с рабочей программой дисциплины по направлению 45.04.02 Лингвистика, магистерская программа – Современные методы прикладной лингвистики и перевода.

Практические задания разработаны в соответствии с рабочей программой дисциплины «Корпусная лингвистика», целью которой является систематическое изложение основных подходов к разработке, сопровождению и алгоритмическому обеспечению электронных корпусов текстов.

Целью практических занятий является закрепление теоретических знаний и приобретение практических умений и навыков, необходимых для осуществления практической деятельности по автоматизированной обработке текстовых массивов с применением электронных корпусов текстов.

Методические рекомендации по каждой практической работе имеют теоретическую часть, необходимую для выполнения практических заданий. Практические задания органично сочетаются с теоретическими знаниями.

Практическое занятие 1**.**

Тема: Проектирование электронных корпусов текстов

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корпусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Проект любого корпуса должен предусматривать этапы его создания и пути его дальнейшего развития. Понятие корпуса является продолжением традиционных картотек, с которыми всегда работали лингвисты. В XX веке эти картотеки стали компьютерными и общедоступными. Значительную роль в становлении корпусного подхода сыграла сеть Интернет, в процессе развития которой стали доступны большие объемы текстового материала, пригодного для проведения различных лингвистических исследований. При этом встает традиционный вопрос о репрезентативности и сбалансированности языкового материала (см. п. 1.4.1), который кладется в основу словарей и грамматик. Особенно остро этот вопрос встает при формировании национальных корпусов. Репрезентативность корпуса должна обеспечиваться как достаточным объемом текстового материала, так и его разнообразием.

Помимо жанрово-тематической структуры предстоит решить также множество других, частных, но важных вопросов, таких как:

1.Что является текстом в корпусе? Например, небольшие объявления в газетах – включаются ли они в корпус как отдельные тексты или их можно объединять?

2.Является ли текстом статья в газете? Или один выпуск газеты нужно расценивать как один текст?

3.Что является отдельным текстом – сборник стихотворений или каждое стихотворение?

4.Является ли отдельным текстом каждое письмо в опубликованной переписке, где авторами писем являются двое, но письма образуют единый дискурс, или совокупность

этих писем?

Не менее важна и проблема хронологии. Что следует понимать под корпусом *современного* русского языка? Представляется, что хронологические рамки корпуса должны быть разными для разных жанров.

Корпус создается для широкого круга пользователей и для решения разнообразных задач, в том числе и достаточно «экзотических», например, для исследования русскоязычных текстов, использующих иноязычную графику. Что из исходных текстов остается в корпусе, а что «вычищается»? Очевидно, например, что картинки не относятся к языковому материалу и могут быть удалены. Сложнее обстоит дело с таблицами и, тем более, с цитатами, прямой речью, иноязычными вкраплениями, единицами измерения.

Все эти вопросы должны быть поставлены на этапе проектирования. Решаться же они, по крайней мере, некоторые из них, могут постепенно в процессе создания и опытной эксплуатации корпуса. Для этого с самого начала эксплуатации следует предусмотреть обратную связь с пользователями.

Практическая часть (вопросы и задания для собеседования)

Расскажите о следующих этапах создания корпусов текстов:

1. Обеспечение поступления текстов в соответствии с перечнем источников.
2. Преобразование в машиночитаемую форму. Тексты в электронном виде для создания корпусов могут быть получены самыми разными способами – ручной ввод, сканирование, авторские копии, дары и обмен, Интернет, оригинал-макеты, предоставляемые издательствами составителям корпусов и др.
3. Анализ и предварительная обработка текстов. На этом этапе все тексты, полученные из разных источников, проходят филологическую выверку и корректировку. Подготовка «технологического» описания включает в себя библиографическое и экстралингвистическое описания текста.
4. Конвертирование и графематический анализ. Некоторые тексты проходят также через один или несколько этапов предварительной машинной обработки, в ходе которых осуществляется перекодировка (если требуется), а также удаление или преобразование нетекстовых элементов (рисунки, таблицы), удаление из текста переносов, «жестких концов строк» (тексты из MS-DOS), обеспечение единообразного написания тире и т.д. Графематический анализ предполагает проведение следующих операций: разделение входного текста на элементы (слова, разделители и т.д.), удаление нетекстовых элементов, выделение и оформление нестандартных (нелексических) элементов, обработка специальных текстовых элементов (имен (имя, отчество), написанных инициалами, иностранных лексем, записанных латиницей, названий рисунков, примечаний, страниц форзаца, зачеркиваний, титульных листов, списков литературы и т.д.). Как правило, эти операции выполняются в автоматическом режиме. Обычно на этом же этапе осуществляется сегментирование текста на его структурные составляющие.
5. Разметка текста. Разметка текста заключается в приписывании текстам и их компонентам дополнительной информации (метаданных). Метаданные можно поделить

на 3 типа: экстралингвистические, относящиеся ко всему тексту; данные о структуре текста; лингвистические метаданные, описывающие элементы текста. Метаописание текстов корпуса включает как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ). Эти данные обычно вводятся вручную. Структурная разметка документа (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически.

6. Корректировка результатов автоматической разметки: исправление ошибок и снятие неоднозначности (вручную или полуавтоматически).

7. Конвертирование размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager), обеспечивающей быстрый многоаспектный поиск и статистическую обработку (заключительный этап).

8. Обеспечение доступа к корпусу. Корпус может быть доступен в пределах дисплейного класса, может распространяться на компакт-диске и может быть доступен в режиме глобальной сети. Различным категориям пользователей могут предоставляться разные права и разные возможности.

9. Создание документационного обеспечения, в котором описываются различные аспекты создания и использования корпуса, в частности, приводятся сведения о разметке, позволяющие искать по метаданным, язык запросов корпус-менеджера и т.д.

Список рекомендуемой литературы

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/

2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753

3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.

4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.

6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.

2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. Vi Improved - http://www.vim.org

Практическое занятие 2**.**

Тема: Отбор источников для электронного корпуса.

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Основной единицей корпуса текстов могут быть *словоупотребления* (обычно их называют словами), *основы* (корни, леммы) и *предложения.* Объем создаваемого корпуса текстов в принятых единицах зависит от целей создания. Он может быть небольшим при изучении частоты употребления букв, буквосочетаний, звуков, звукосочетаний. Гораздо бóльшим он должен быть при изучении лексики, морфологических явлений и при изучении синтаксических или стилистических особенностей текстов [17]. Проблемными являются также следующие вопросы:

7. Тексты каких функциональных жанров включать в корпус текстов (художественную прозу, драму, стихи, научные тексты, газеты, журналы, технические описания и т.д.)?
8. Тексты каких временны*х* промежутков включать в корпус текстов (современные, 10-летней давности, 50-летней давности, древние и т.д.)?
9. Включать ли тексты только литературного языка или также другие типы источников? И что считать литературным языком?

При ответе на эти вопросы разработчики корпуса текстов обычно используют консультации специалистов по языкознанию и лингвостатистике или метод анкет. Исходя из своего опыта исследований, специалисты определяют общий объем корпуса текстов, время издания текстов, число текстов и размер элементарной выборки, жанры отбираемых текстов и их количество, число элементарных выборок из каждого жанра. Метод анкет в сочетании с опытом специалистов был использован при создании корпуса текстов «Аме-

риканский корпус наследия» (TheAmericanHeritageIntermediateCorpus). Специалисты определили его объем в 5 млн. слов (словоупотреблений) и рекомендовали включить в него лексику из 22 разделов (жанров) детской и юношеской литературы на английском языке. В 221 школу США были разосланы анкеты с просьбой указать, какие тексты желательно включить в корпус. После изучения анкет был составлен список из 19 тыс. названий книг. Из этого множества было отобрано 1045 текстов. На их основе было составлено 10 тыс. элементарных выборок по 500 словоупотреблений каждая [17].

Практическая часть (вопросы и задания для собеседования)

1. Что должно являться основной единицей корпуса текстов?
2. Каким должен быть объем корпуса текстов (сколько единиц он должен содержать)?
3. Какие письменные текстовые источники должны быть представлены в корпусе текстов и в каком количестве?
4. Из какой исходной языковой области должны быть выбраны тексты, включаемые в состав корпуса?

Список рекомендуемой литературы

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. **Vi Improved - http://www.vim.org**

Практическое занятие 3.

Тема: **Основные процедуры обработки естественного языка: токенизация, лемматизация, стемминг, парсинг**

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Фактически, корпус в его современном понимании – это всегда компьютерная база данных, и в процессе его создания естественно использование специальных процедур и программ. Например, ***токенизация***, т.е. разбиение потока символов в естественном языке на отдельные значимые единицы (токены, словоформы), является необходимым условием для дальнейшей обработки естественного языка. Если бы языки обладали совершенной пунктуацией, токенизация не представляла бы сложности – даже самая простая программа могла бы разделить текст на слова, руководствуясь пробелами и знаками препинания. Но в действительности языки подобной пунктуацией не обладают, что усложняет задачу токенизации. Например, в английском языке встречаются случаи, которые не могут быть однозначно токенизированы. Ср.: строка *chap*. может являться сокращенной формой слова *chapter* или словом *chap*, которое расположено в конце предложения. Строку *Jan*. можно рассматривать как сокращенную форму слова *January* либо как имя собственное, расположенное в конце предложения. В первом случае точка должна быть отнесена к тому же токену, что и слово, а во втором случае она должна быть выделена в отдельный тэг. Вместе с тем, нельзя не заметить, что подобные трудности весьма ограничены, и многие приложения, обрабатывающие текст, часто игнорируют их (например, не учитывают аббревиатуры и сложные слова), либо обрабатывают их с помощью отдельного алгоритма.

Другая специфическая задача морфологического анализа – это ***лемматизация***, т.е.

процесс образования первоначальной формы слова, исходя из других его словоформ. Во многих языках слово может встречаться в нескольких формах с различными флексиями. Например, английский глагол *'walk'* может быть представлен следующими формами: *'walk', 'walked', 'walks', 'walking'*. Базовая форма, *'walk'*, зафиксированная в словаре, называется ***леммой*** слова. Лемматизация представляет собой процесс группировки различных флективных форм одного слова таким образом, чтобы при анализе они обрабатывались как одно слово.

Процесс, несколько отличный от лемматизации, называется ***стеммингом***, он состоит в нахождении стема (основы) слова. Разница заключается в том, что стеммер обрабатывает отдельное слово без знания контекста, и, таким образом, не может дифференцировать слова, которые имеют разные значения в силу отнесенности к разным частям речи. Тем не менее, стеммеры обычно более просты для реализации и быстрее обрабатывают данные, а более низкая точность их работы может не иметь решающего значения для многих приложений. Например, токену "*better*" соответствует лемма "*good*", но это опускается при стемминге. Лемма "*walk*" является базовой формой для токена "*walking*", и это соответствие будет обнаружено как при стемминге, так и при лемматизации.

Ниже приведены примеры стемминга и лемматизации. Данослеующееепредложение:

[The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dogs].

Один из наиболее популярных стеммеров, SnowballAnalyzer, выдает следующие стемы:

[quick] [brown] [fox] [jump] [over] [lazy] [dog].

Леммы слов данного предложения будут следующими:

[the] [quick] [brown] [fox] [jump] [over] [the] [lazy] [dog].

Лемматизация связана с идентификацией частей речи и включает в себя сокращение слов из корпуса до соответствующих им лексем. Именно лемматизация позволяет исследователю выделять и изучать все варианты отдельной лексемы без необходимости введения всех возможных вариантов. Рассмотрим пример работы морфологического анализатора с английским предложением *"Allwomenwerewalkinginthestreets"*. Токены (словоформы) представлены слева в скобках <>, звездочка '*' показывает, что слово в тексте начинается с заглавной буквы. Под каждым токеном располагается лемма (лексема) и приводится морфологический разбор. Например, токен "were" относится к лемме "be", и его морфологические характеристики – глагол, прошедшее время, спрягаемый; токен "streets" относится к лемме "street", и его морфологические характеристики – существительное, нарицательное, ед. числа и т.д.

```
"<*all>"
    "all" <*><Quant> DET PRE SG/PL
"<women>"
    "woman" N NOM PL
"<were>"
```

"be" <SV><SVC/N><SVC/A> V PAST VFIN
"<walking>"
"walk" <SV><SVO> PCP1
"<in>"
"in" PREP
"<the>"
"the" <Def> DET CENTRAL ART SG/PL
"<streets>"
"street" N NOM PL
"<$.>"

Парсинг – это процесс сопоставления линейной последовательности лексем (слов, токенов) языка с его формальной грамматикой. Результатом обычно является дерево зависимостей (синтаксическое дерево). Построение автоматических синтаксических анализаторов (парсеров) для больших корпусов является одной из самых важных областей компьютерной лингвистики. Большинство подходов объединяет качественные и количественные измерения. Наряду с разными статистическими подходами, которые тренируются на снабженных вручную пометами синтаксических деревьях (tree-banks), многие синтаксические анализаторы используют основанные на правилах или основанные на ограничениях подходы, которые прямо моделируют специфические лингвистические теории. Разработка этих синтаксических анализаторов тесно переплетается с развитием этих теорий. Поскольку большинство предложений неоднозначны в любой теории, на основе правил (или перечня ограничений) должна быть разработана стратегия снятия неоднозначности. Многие стратегии снятия неоднозначности полагаются на количественные данные – частоту данной структуры в данном корпусе (тип), ограничения на выборку для данных лексических единиц, которые были получены или выделены из корпусных данных, и т.д.

Необходимо рассматривать два условия при обсуждении предварительной обработки корпусов:

1. Каждый шаг подготовки текста к обработке заставляет составителя корпуса принимать лингвистические решения, которые влияют на последующие шаги и на оценку корпуса. Конечный пользователь должен быть в курсе этих решений, чтобы найти то, что он ищет. Например, тот, кто делит тексты на составные части, должен решить, относиться к случаям типа *New York* и *Baden Baden* как к одному слову или как к двум. Подобным образом, человек, выявляющий лексемы, должен решить, что делать с такими явлениями, как немецкие глаголы с отделяемыми приставками.

2. Конечного пользователя нужно поставить в известность о том, какая работа была проделана на стадии предварительной обработки и о возможных погрешностях, поскольку любые ошибки в кодировке, особенно системные, могут повлиять на результаты, полученные пользователями корпуса [42].

Практическая часть (вопросы и задания для собеседования)

Охарактеризуйте и продемонстрируйте на примере ПО GATE каждый из следующих этапов обработки текста:

а) токенизация;

б) лемматизация;

в) стемминг;

г) парсинг

Список рекомендуемой литературы

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs

6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. **Vi Improved - http://www.vim.org**

Практическое занятие 4.

Тема: **Разметка. Средства разметки корпусов.**

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Среди специальных программ для обработки естественного языка особое место занимают программы автоматической разметки. Разметка корпусов (tagging, annotation) представляет собой трудоемкую операцию, особенно учитывая размеры современных корпусов. Если для некоторых видов разметки, в частности, анафорической, просодической, создание автоматических систем пока представляется довольно сложным и основная часть работы проводится вручную, то для морфологического и синтаксического анализа существуют различные программные средства, которые принято называть соответственно тэггеры (taggers) и парсеры (parsers). В результате работы программ автоматического морфологического анализа (тэггеров) каждой лексической единице приписываются грамматические характеристики, включая часть речи, лемму и набор граммем (например, род, число, падеж, одушевленность/неодушевленность, переходность и т.д.). В результате работы программ автоматического синтаксического анализа фиксируются синтаксические связи между словами и словосочетаниями, а синтаксическим единицам приписываются соответствующие характеристики (тип предложения, синтаксическая функция словосочетания и т.д.).

Однако автоматический анализ естественного языка небезошибочен и многозначен – он, как правило, дает несколько вариантов анализа для одной лексической единицы (слова, словосочетания, предложения). В этом случае говорят о грамматической омонимии. Снятие неоднозначности (морфологической, синтаксической) в целом является одной из важнейших и сложнейших задач компьютерной лингвистики. При создании

корпусов для снятия неоднозначности используются автоматические и ручные способы. Корпусы нового поколения включают сотни миллионов слов, поэтому выдвигаются принципы разработки систем, которые бы минимизировали вмешательство человека. Автоматическое разрешение морфологической или синтаксической неоднозначности, как правило, основывается на использовании информации более высокого уровня (синтаксического, семантического) с применением статистических методов.

Для решения различных лингвистических задач недостаточно иметь массив текстов. Требуется также, чтобы тексты содержали в себе явным образом указанную разного рода дополнительную лингвистическую и экстралингвистическую информацию. Так, на материале корпуса, подобного Брауновскому, можно легко выявить ***частотность*** слов – их регулярное употребление в определенных контекстах. Однако это будет частотность токенов (словоформ). Для определения частоты лексем каждому слову должна быть приписана ее лемма.

Для подсчета частот в разрезе грамматических категорий они также должны быть маркированы. В корпусе, снабженном такой информацией, существительные имеют, например, тэг *noun*, глаголы – тэг *verb* и т.д. Помимо прочего, такие тэги позволяют изучать групповые характеристики слов, имеющих определенную помету. Если снабжать тэгами слова в большом корпусе вручную, это займет очень много времени, поэтому исследователи разработали способы автоматической разметки в корпусе. Один из простых способов заключается в том, чтобы компьютеризированный словарь, в котором указаны лексические категории для самых распространенных слов или для наибольшего количества слов, совместить с неразмеченным корпусом. Затем каждому слову в неразмеченном корпусе может быть автоматически присвоен тэг от соответствующего ему слова в снабженном пометами словаре. Таким образом, если словоформы *information* и *distribution* появились и в корпусе, и в словаре, тэг ‘noun’, который сопровождал эти словоформы в словаре, автоматически будет перенесен на них в корпусе. Подобно этому, такие формы как *lexical* и *frequent* будут помечены как прилагательные, поскольку они всегда являются членами этой категории, *the* и *a* будут помечены как артикли, *identify* и *see* – как глаголы и т.д. [42].

Этот процесс нахождения соответствующих форм в корпусе и в снабженном пометами словаре не может быть использован для определения категорий всех форм, потому что некоторые формы могут быть членами более чем одной категории. Эта проблема носит название «проблема морфологической неоднозначности (ambiguity)». Например, слова *words, forms, can, use, present* и *process* могут быть как существительными, так и глаголами. Поскольку в английском языке так много форм принадлежит более чем одной категории, точно разметить слова можно благодаря более сложным процедурам, чем автоматическое совмещение со словарем. Конечно, в контексте (т.е. в действительном использовании) словоформа принадлежит только одной категории. Следовательно, достичь точной разметки английского корпуса можно путем анализа контекста или анализа более высокого уровня: синтаксического анализа для морфологической разметки, семантического – для синтаксической.

Возьмем слово *deal* в качестве примера. Как словоформа, оно может быть как

существительным, так и глаголом. Предположим, что корпус содержал фразу *a good deal of trouble*, и предположим, что автоматическое совмещение со словарем уже позволило пометить *good* как прилагательное. При выборе между тем, предшествует ли прилагательное существительному или глаголу, намного надежнее выбрать существительное, поскольку в английском языке прилагательные обычно предшествуют существительным и обычно не предшествуют глаголам. Так, *deal* в *a good deal of trouble* может быть помечено как существительное. Другими словами, поскольку *good* однозначно является прилагательным, оно будет помечено как adjective на начальном уровне снабжения пометами путем совмещения корпуса со словарем. Если начинать разметку, размечая только слова, принадлежащие исключительно одной категории, а затем использовать эту информацию для того, чтобы прояснить неоднозначные случаи, многие сложные проблемы смогут быть решены. В обычной практике случается так, что слова снабжаются пометами сначала для всех частей речи, к которым они могут относиться, а затем категории примыкающих слов используются для определения категории слов, у которых есть несколько помет.

Итак, *разметка* заключается в приписывании текстам и их компонентам специальных тэгов: собственно *лингвистических*, описывающих лексические, грамматические и прочие характеристики элементов текста, и внешних, *экстралингвистических* (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика).

Практическая часть (вопросы и задания для собеседования)

1) Понятие аннотирования.

2) Ручное и автоматическое аннотирование текста.

3) Принципы аннотирования текста в GATE.

Список рекомендуемой литературы

Основная литература:
1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:
1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.

5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. **Vi Improved - http://www.vim.org**

Практическое занятие 5.

Тема: Лингвистическая разметка

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Среди лингвистических типов разметки выделяются:морфологическая, синтаксическая, семантическая, анафорическая, просодическая, дискурсная и др.

Морфологическая разметка

В иностранной терминологии употребляется термин part-of-speech tagging (POS-tagging), дословно – частеречная разметка. В действительности морфологические метки включают не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. Это основной тип разметки: во-первых, большинство крупных корпусов являются как раз морфологически размеченными корпусами, во-вторых, морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического, и, в-третьих, успехи в компьютерной морфологии позволяют автоматически с большой степенью правильности размечать корпусы больших размеров.

Данные о разметке представляются в том или ином структурированном виде и включают: лемму, признак части речи, признаки грамматических категорий. В 1980 году появилась размеченная версия Брауновского корпуса, в которой была проведена лемматизация словоформ, маркировка их поверхностно-синтаксических функций и т.д. Морфологическая разметка Брауновского корпуса выглядит следующим образом:

the_AT jury_NN further_RB said_VBD in_IN term-end_NN presentments_NNS that_CS the_AT *city_NP *executive_NP *committee_NP ,_, which_WDT had_HVD over-all_JJ charge_NN of_IN the_AT election_NN ,_, deserves_VBZ the_AT praise_NN and_CC thanks_NNS of_IN the_AT *city_NP of_NP *atlanta_NP for_IN the_AT manner_NN in_IN which_WDT the_AT election_NN was_BEDZ conducted_VBN |

Приведем пример морфологической разметки фрагмента текста на русском языке «*Звонили к вечерне. Торжественный гул колоколов*» в XML-формате на основе разметчика АОТ (*рис. 1*).

В представленной записи использованы тэги <text> – текст, <p> – абзац, <s> – предложение, <w> – словоупотребление, <pun> – знак пунктуации. Тэг <w> содержит вложенный тэг <ana> с атрибутами <lemma> – лемма, <pos> – часть речи, <gram> – набор граммем. Значения граммем приводятся в Приложении 3.

Синтаксическая разметка

Синтаксическая разметка является результатом парсинга, выполняемого на основе данных морфологического анализа. Этот вид разметки описывает синтаксические связимежду лексическими единицами и различные синтаксические конструкции (например, придаточное предложение, глагольное словосочетание и т.д.).

```
<?xml version="1.0" encoding="windows-1251" ?><text><p>
<s>
<w>Звонили<ana lemma="ЗВОНИТЬ" pos="Г" gram="мн,нс,нп,дст,прш," /></w>
<w>к<ana lemma="К" pos="ПРЕДЛ" gram="" /></w>
<w>вечерне
<ana lemma="ВЕЧЕРНЯ" pos="С" gram="жр,ед,дт,пр,но," />
<ana lemma="ВЕЧЕРНИЙ" pos="П" gram="ср,ед,кр," /></w>
<pun>.</pun></s>
<s><w>Торжественный<ana lemma="ТОРЖЕСТВЕННЫЙ" pos="П"
gram="мр,ед,им,вн," /></w>
<w>гул<ana lemma="ГУЛ" pos="С" gram="мр,ед,им,вн,но," /></w>
<w>колоколов
<ana lemma="КОЛОКОЛ" pos="С" gram="мр,мн,рд,но," />
<ana lemma="КОЛОКОЛОВ" pos="С" gram="мр,фам,ед,им,од," /></w>
……..<pun>.</pun></s></p></text>
```

Рис. 1. Пример морфологической разметки текста на русском языке (список граммем см. Приложение 3)

В отличие от морфологии, способы представления синтаксической структуры и синтаксических отношений не столь унифицированы. Наблюдается разнообразие синтаксических теорий и формализмов:

- грамматика зависимостей;

- грамматика непосредственно составляющих;
- грамматика структурных схем;
- традиционные синтаксические учения о членах предложения;
- функциональная грамматика;
- семантический синтаксис и др.

Синтаксический анализ для русского языка чаще всего представлен структурами зависимостей. Нарисунке 2 представленпримервизуализациидеревазависимостей.

Long ago, in the city of Babylon, the people began to build a huge tower which seemed to reach the heavens soon.

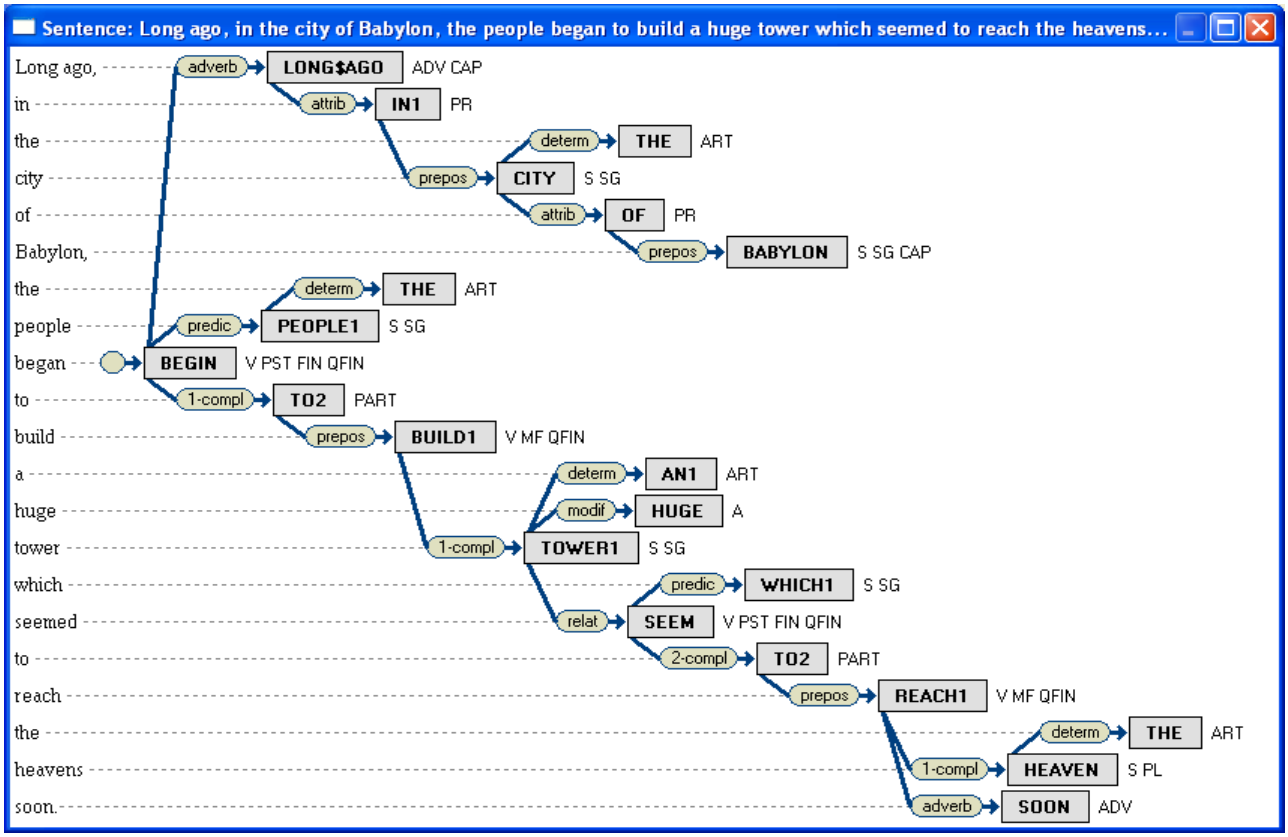


Рис. 2. Пример синтаксического разбора
(грамматика зависимостей, система ЭТАП-3)

Семантическая разметка

Семантические тэги чаще всего обозначают семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, специфицирующие его значение. Семантическая разметка корпусов предусматривает спецификацию значения слов, разрешение омонимии и синонимии, категоризацию слов (разряды), выделение тематических классов, признаков каузативности, оценочных и деривационных характеристик и т.д.

Свой вариант семантической разметки предлагает НКРЯ. В этом корпусе каждой словоформе приписываются пометы трех типов.

1) разряд (имя собственное, возвратное местоимение и т.д.);

2) лексико-семантические характеристики (тематический класс лексемы, признаки каузативности, оценки и т.д.);

3) деривационные характеристики («диминутив», «отадъективное наречие» и т.д.).

Собственно лексико-семантические тэги сгруппированы по следующим полям:

- таксономия (тематический класс лексемы) – для имен существительных, прилагательных, глаголов и наречий;
- мереология (указание на отношения «часть – целое», «элемент – множество») – для предметных и непредметных имен;
- топология (топологический статус обозначаемого объекта) – для предметных имен;
- каузация – для глаголов;
- служебный статус – для глаголов;
- оценка – для предметных и непредметных имен, прилагательных и наречий.

Словообразовательные характеристики включают несколько типов:

- морфо-семантические словообразовательные признаки (например, «каритив», «семельфактив»);
- разряд производящего слова (например, отглагольное существительное или отадъективное наречие);
- лексико-семантический (таксономический) тип производящего слова (например, наречие, образованное от прилагательного размера);
- морфологический тип словообразования (субстантивация, сложное слово) (более подробно см. http://ruscorpora.ru, раздел «Семантика»).

Существуют и другие типы разметки, в частности:

- *анафорическая* разметка. Она фиксирует референтные связи, например, местоименные;
- *просодическая* разметка. В просодических корпусах применяются тэги, обозначающие ударение и интонацию. В корпусах устной разговорной речи просодическая разметка часто сопровождается так называемой *дискурсной* разметкой, которая служит для обозначения пауз, повторов, оговорок и т.д.

Практическая часть (вопросы и задания для собеседования).

Расскажите о следующих принципах разметки текста:

1) описание (обоснование) схемы разметки;
2) общепринятая система лингвистических понятий;
3) известная для пользователя схема анализа;
4) мотивированность введения параметров;
5) теоретически нейтральная (традиционная) схема разметки;
6) следование международным стандартам.

Список рекомендуемой литературы

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. **Vi Improved - http://www.vim.org**

Практическое занятие 6.

Тема: Экстралингвистическая разметка

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Экстралингвистическая разметка, или метаданные, включает в себя «внешнюю», «интеллектуальную» разметку (библиографические характеристики, типологические характеристики, тематические характеристики, социологические характеристики), «формальную» структурную разметку (текст, раздел, глава, часть, абзац, предложение), а также технико-технологическую разметку (кодировку, даты обработки, исполнителей, источник электронной версии). Набор метаданных во многом определяет возможности, предоставляемые корпусами исследователям. При выборе этих данных необходимо руководствоваться целями исследования и потребностями лингвистов, а также возможностями по внесению в текст тех или иных дополнительных признаков.

«Внешняя», «интеллектуальная» разметка нужна, во-первых, для выявления взаимосвязи языка и условий его существования; во-вторых, для изучения отдельных подмножеств языка. Выделяют два класса факторов, влияющих на язык текстов:

- внешние, внеязыковые факторы (E – external);
- внутренние факторы (I– internal).

Дж. Синклер выделяет три группы ***E-факторов***:

- E1 (origin) – факторы, относящиеся к созданию текста автором;
- E2 (state) – факторы, относящиеся к внешним признакам текста (включая устную или письменную речь);
- E3 (aims) – факторы, относящиеся к причинам создания текста и его влиянию на ауди-

торию

и две группы ***I-факторов***:

- I1 (topic) – предметная область текста;
- I2 (style) – стилистические особенности (стиль, жанр) [57].

В НКРЯ, например, используется следующий набор метаданных:

Первый блок:

1) *автор текста*: имя, пол, дата рождения (или примерный возраст);

2) *название текста*;

3) *время и место создания текста* (может указываться точно или приблизительно);

4) *объем текста*: для художественных произведений принято, что обычная длина рассказа – менее 5 тыс. слов; обычная длина повести – от 5 до 15 тыс. слов; обычная длина романа – более 15 тыс. слов.

Второй блок: параметры метаописания трех основных *массивов* текстов корпуса – художественных текстов; нехудожественных текстов; драматургических произведений. Например, для художественных текстов в НКРЯ указывается:

1) жанр текста: нежанровая проза, автобиографическая проза, детектив, детская литература, историческая проза, криминальная литература, приключения, фантастика, юмор и сатира;

2) тип текста: автобиографическая проза, анекдот, ассоциативная проза, боевик, детектив, очерк, литературное письмо, повесть, притча, пьеса, рассказ, роман, сказка, триллер, эпопея, эссе и др.;

3) хронотоп текста: приблизительное указание на место и время описываемых в тексте событий [27].

Реально предлагается следующее: древний Восток; Россия XVII век; Россия XIX век; Россия/СССР: советский период в целом; Россия, советский период – Германия 1920-1940-е годы; Россия/СССР – Европа 1960-1980-е годы; Россия/СССР: перестройка; Россия/СССР: советский и постсоветский период; Америка: современная жизнь; Израиль: современная жизнь; Средняя Азия: современная жизнь; ирреальный мир и др. Также может встретиться тэг «хронотоп не определен».

Служебная, или «имплицитная», метаразметка в НКРЯ включает:

1) «текст-стиль», при этом выделяются академический, научно-популярный, официально-деловой, нейтральный, сниженный, сниженный с элементами грубого просторечия и жаргона, архаизованный, индивидуально-авторский, диалектный и пр. (всего 21);

2) аудитория – возраст;

3) аудитория – уровень образования;

4) аудитория – размер (более подробно см. http://ruscorpora.ru/corpora-parameter.html)

Практическая часть (вопросы и задания для собеседования)

Расскажите о принципах экстралингвистической разметки в контексте следующих основных факторов:

- E1 (origin) – факторы, относящиеся к созданию текста автором;
- E2 (state) – факторы, относящиеся к внешним признакам текста (включая устную или письменную речь);
- E3 (aims) – факторы, относящиеся к причинам создания текста и его влиянию на аудиторию
- I1 (topic) – предметная область текста;
- I2 (style) – стилистические особенности (стиль, жанр) [57].

1) *автор текста*: имя, пол, дата рождения (или примерный возраст);

2) *название текста*;

3) *время и место создания текста* (может указываться точно или приблизительно);

4) *объем текста*: для художественных произведений принято, что обычная длина рассказа – менее 5 тыс. слов; обычная длина повести – от 5 до 15 тыс. слов; обычная длина романа – более 15 тыс. слов.

Список рекомендуемой литературы

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. **Vi Improved - http://www.vim.org**

Практическое занятие 7.

Тема: Корпус как поисковая система

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Использование находящегося в свободном доступе достаточно большого количества инструментов обработки текста превращает коллекции текстов в электронные продукты, которые могут накапливать и обрабатывать лингвистическую информацию согласно задачам исследователя.

Неотъемлемой частью понятия «корпус текстов» является система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего называют ***корпусным менеджером***(или корпус-менеджером). Корпусный менеджер –это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

Корпусный менеджер должен:

- строить как KWIC (KeyWordInContext), так и полные конкордансные списки;
- искать не только отдельные слова, но и словосочетания;
- осуществлять поиск по шаблонам (сложные запросы);
- сортировать списки по нескольким критериям, выбираемым пользователем;
- давать возможность отображать найденные словоформы в неограниченном контексте;
- давать статистическую информацию по отдельным элементам корпуса;

- отображать леммы, морфологические характеристики словоформ и метаданные (библиографические, типологические), что зависит от степени размеченности корпуса;
- сохранять и распечатывать результаты;
- работать как с отдельными файлами, так и с корпусами, неограниченными по размеру;
- быстро обрабатывать запросы и выдавать результаты;
- поддерживать различные форматы текстовых данных (txt, doc, rtf, html, xml и др.);
- быть легким (интуитивно понятным) в использовании, как для опытного, так и для начинающего пользователя.

Наиболее известны такие универсальные корпусные менеджеры как SARA, XAIRA (BNC), Manatee/Bonito, CQP, DDC. Для обработки корпусных данных могут разрабатываться менеджеры на основе системуправлениябазамиданных (СУБД) или поисковых систем.

Например, поиск по Национальному корпусу русского языка осуществляется поисковой системой Яndex.Server 3.8 Professional. Для поиска грамматической и метатекстовой информации задействованы способности Яndex.Server по поиску скрытых свойств (атрибутов) документов и фрагментов текста. Поисковая выдача формируется при помощи средств Яndex.Server, который обеспечивает полнотекстовый поиск информации с учетом морфологии русского языка на веб-сервере или в корпоративной сети. Поиск работает с учетом морфологии русского, английского и украинского языков – так же, как работает поиск Яндекс по Интернету. Например, если задан запрос «идти», то в результате поиска будут найдены ссылки на документы, содержащие слова «идти», «идет», «шел», «шла» и т.д. Результатом поиска является список документов, упорядоченных по релевантности, которая учитывает не только количество найденных документов, но и контрастность слов (частоту их употребления) и расстояние между словами [27].

3.1.2. Языки запросов

Информационный запрос – это словесное выражение определенной информационной потребности. Запросы анализируются по своему предметному и формальному содержанию и описываются в терминах словаря языка запросов прикладной программы, работающей с корпусом. Процедура поиска заключается в поочередном сопоставлении поискового образа запроса с отдельными элементами корпуса и в вычислении их соответствия. При наличии такого соответствия элементы корпуса текстов считаются релевантными и подлежат выдаче.

В общем виде модель языка запросов включает в себя следующие элементы:

1) собственно поисковые элементы (термины, выражающие информационную потребность и т.д.);

2) средства морфологической нормализации текстовых элементов запроса;

3) булевы операторы (конъюнкция, дизъюнкция, отрицание);

4) средства линейной грамматики (операторы расстояния, позиционные операторы);

5) дополнительные условия поиска:

- поиск в определенных полях корпуса (например, внутри тэгов);
- ограничение области поиска (по произведениям определенных авторов, по дате создания документов, их типу и т.д.);

6) требование на сортировку (ранжирование) выдаваемых результатов;

7) требования к форме представления результатов поиска:

- вид выдаваемых результатов;
- количество выдаваемых документов.

Далее будет рассмотрен язык запросов одного из наиболее эффективных корпусных менеджеров, **Bonito/Manatee**[1]. На примере этой поисковой системы будет продемонстрировано большинство основных элементов языка запросов к корпусам текстов, а также приведены примеры задания запросов к корпусу.

Корпусный менеджер Bonito представляет собой программное обеспечение для работы с корпусами текстов. Система Bonito состоит из двух частей: сервера (Bonitosrv) и графического пользовательского интерфейса (GUI – graphical user interface) Bonito, созданного П. Рыхли и группой NLPlab (Natural Language Processing Laboratory) на факультете информатики Университета им. Масарика (Чехия) и работающего на стороне клиента.

Для демонстрации работы с системой будет использоваться корпус английских текстов SUSANNE (SurfaceandUnderlyingStructuralAnalysisofNaturalEnglish) (http://www.grsampson.net/). Данный корпус был создан в Великобритании в Университете Сассекса. Он включает в себя более 130 тыс. слов Брауновского корпуса американского английского языка, аннотированного согласно схеме SUSANNE.

Основные особенности системы Bonito

Язык запросов

- поиск отдельных атрибутов (словоформа, лемма, тэг);
- использование регулярных выражений;
- логические операторы;
- средства задания структуры (границы предложения и др.);
- быстрая обработка сложных запросов;
- шаблоны.

Конкордансные списки

- история запросов пользователя;
- просмотр морфологических характеристик словоформы;

---

[1] Bonito – название менеджера, Manatee – вся программная подсистема корпусного обеспечения.

- отображение леммы.

Операции над конкордансом

- сохранение списков в файл;
- печать списков;
- сортировка по ключевым словам, контексту;
- интерактивное неограниченное расширение контекста;
- фильтрация (удаление части построенных конкордансов);
- удаление повторений.

Частотное распределение

- частоты слов и других атрибутов в корпусе, контексте;
- неограниченное число уровней группировки.

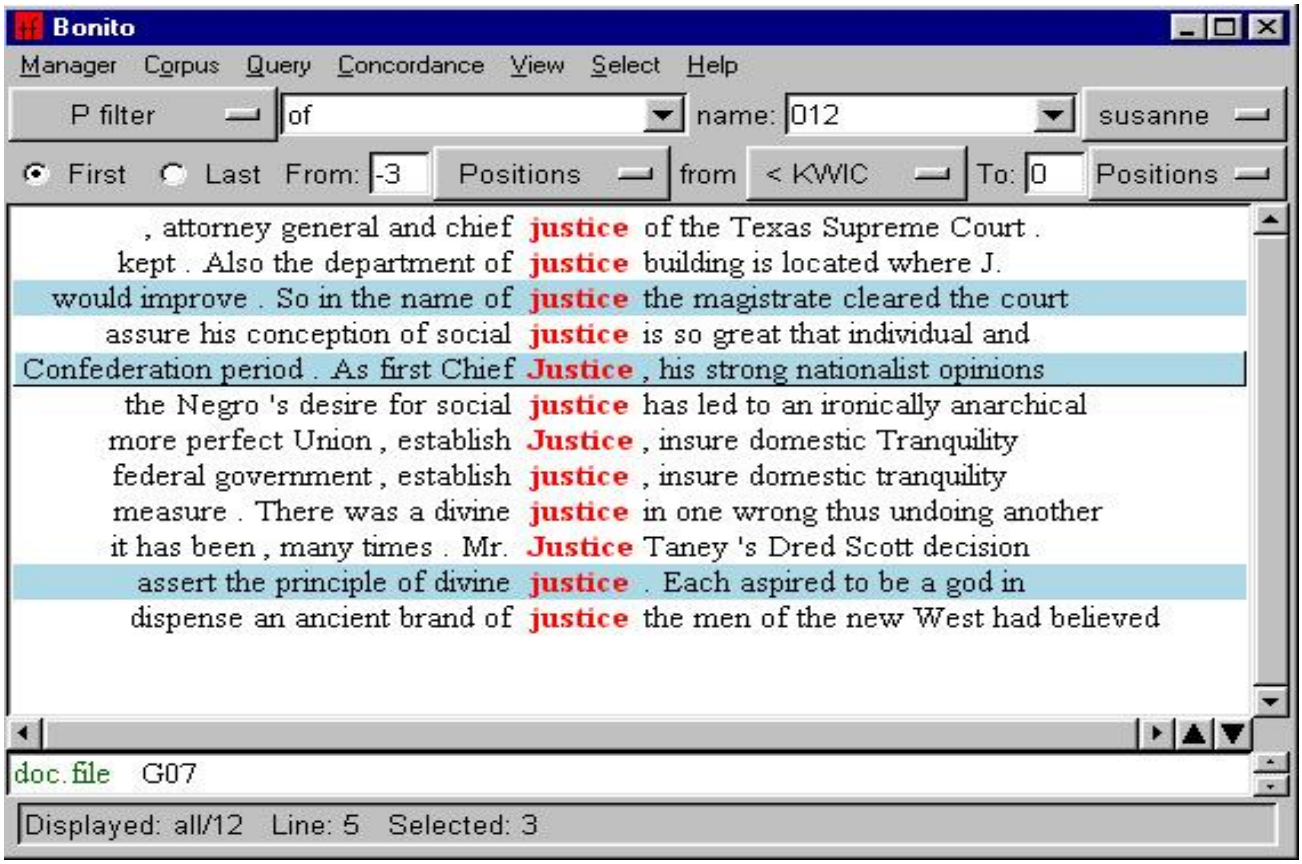
Другие особенности

- выбор кодировок;
- создание пользовательских подкорпусов;
- произвольный набор тэгов;
- возможность подключения других языков.

Запросы

Пользователь может ввести собственно запрос, сформулированный по правилам языка запросов системы, или шаблон (готовый или созданный пользователем) в окно запросов (*рис. 3*).

Рис. 3.Окно корпус-менеджера Bonito с конкордансом для словоформы"justice"



Типы запросов:

Положительный фильтр (P-filter) – совпадающие с запросом строки выдаются в конкордансном списке;

Отрицательный фильтр (N-filter) – совпадающие с запросом строки удаляются из конкордансного списка;

Словосочетания (Collocations) – удовлетворяющие запросу позиции (конкретная словоформа на заданном интервале) в конкордансе выделяются цветом.

Для положительного, отрицательного фильтров и словосочетаний необходимо задавать интервал, в пределах которого следует искать совпадающие позиции для каждой строки конкорданса. Пользователь задает границы интервала (окна ввода "**From:**" и "**To:**"). Если значения положительные, то поиск организуется вправо от исходной позиции, если отрицательные – то влево. Исходной позицией может служить начало словоформы, конец словоформы, начало N-ой позиции, конец N-ой позиции. Очень важно отметить, что все введенные запросы сохраняются в так называемой Истории запросов (Query History), но если запрос идентичен одному из предыдущих, он не попадает в Историю запросов. Достаточно нажать стрелку "вниз" в окне запроса, чтобы проследить всю Историю, а если необходимо, то вернуться к одному из предыдущих введенных запросов.

Если ввести имя запроса в окне "**name:**",запрос сохраняется в списке "названных" (проименованных) запросов (named queries).

Шаблоны

Шаблон – это вид запроса, который упрощает ввод однотипных запросов. Это означает, что сложный запрос необходимо создать только один раз и сохранить как шаблон, а затем просто вводить значения для данного шаблона.

Например, шаблон для всех словоформ правильного английского глагола "play" мог бы выглядеть так:

[word="$1" | word="$1s" | word="$1ed" | word="$1ing"]

В этом шаблоне использовалась переменная, состоящая из значка "$" и цифры "1". Количество переменных в шаблоне не ограничено. При использовании шаблона первый вводимый параметр соответствует переменной $1, второй – $2 и т.д. Параметры вводятся через пробел.

Когда шаблон активизируется, он автоматически записывается в окно запроса. Отличие от обычного запроса состоит лишь в следующем: первый знак строки – это восклицательный знак (**!**), далее идет имя шаблона, двоеточие (**:**) и параметры, разделяемые пробелами. Если бы имя приведенного выше шаблона было "regular verb", то строка запроса для всех форм глагола "play" выглядела бы так:

!regular verb: play

Язык регулярных выражений RegEx

Языки запросов корпусных менеджеров, представленные в той или иной форме (формализованный язык запросов или оконный интерфейс), как правило, базируются на

формализме, который получил название «язык регулярных выражений». Бóльшую часть запросов на языке RegEx «скрывают» от пользователя в программном коде, реализовав их в виде удобного интерфейса. Пользователю необходимо лишь заполнить определенные поля формы (web-страница с ячейками для заполнения), и его запрос будет осуществлен. Но все же для сложных запросов полезно знать основы языка регулярных выражений.

Иногда приходится сталкиваться с ситуацией, когда в операционной системе необходимо найти все файлы с заданным расширением. Известно, что нужно вызвать функцию поиска файлов и в поле "Имя" ввести: *.jpg, тем самым сообщая поисковой машине, что нужно найти файлы, имя которых состоит из любого количества любых символов (*), а расширение должно быть "jpg". В данном примере показано использование регулярных выражений.

Регулярные выражения – это строковые записи, задающие правила поиска на особом языке. Если есть выражение и какая-либо строка (слово, массив текстов, записи в полях базы данных и т.д.), то операцию проверки, удовлетворяет ли строка выражению, называют *сопоставлением* строки и выражения. Если какая-то строка или часть строки успешно сопоставилась с выражением, это называется *совпадением* (соответствием). Например, при сопоставлении выражения "группа букв, окруженная пробелами" и строки "*помню чудное мгновенье*" совпадением будет строка "*чудное*" (ведь только она удовлетворяет данному выражению).

Существует несколько разновидностей языков, используемых для записи регулярных выражений и работы с ними. У них есть много общего, но отдельные части все же отличаются. В популярном языке программирования PHP4 и СУБД MySQL реализован язык регулярных выражений RegEx.

В языке RegEx каждое выражение состоит из одной или нескольких управляющих команд. Некоторые из них можно группировать, и тогда они принимаются за одну команду. Все управляющие команды разбиваются на три класса:

1) *простые символы*, а также *управляющие символы*, играющие роль их заменителей;

2) *управляющие конструкции* (квантификаторы повторений, оператор альтернативы, группирующие скобки и т.д.);

3) так называемые *мнимые символы* (в строке их нет, но они "помечают" какую-то часть строки – например, ее конец).

Простые символы

Класс простых символов, действительно, самый простой. А именно, любой символ в строке на языке RegEx обозначает сам себя, если он не является управляющим. К управляющим символам причисляются следующие: .*?+[]{}|$^

Например, регулярное выражение "abcd" будет "реагировать" на строки, в которых встретится последовательность "abcd".

Группы символов

Одним из самых важных управляющих символов является точка "**.**", обозначающая один любой символ. Например, выражение "л.к" имеет совпадение для строк "лик", "лук",

"лак". Позже будет показано, как можно с помощью точки обозначить ровно один любой символ (или, к примеру, ровно пять).

Возможно, понадобится искать не любой символ, а один из нескольких указанных. Для этого нужно заключить их в квадратные скобки. К примеру, выражение "л[иуа]к" соответствует строкам, в которых есть подстроки из трех символов, начинающиеся с "л", затем одной из букв "и,у,а" и, наконец, "к". Если букв-альтернатив много, и они идут подряд (в алфавитном порядке), то не обязательно перечислять их все. Достаточно указать через дефис первую и последнюю. Например, выражение "[а-я]" обозначает любую букву от "а" до "я", а выражение "[а-я0-9]" задает любой алфавитно-цифровой символ.

Существует и другой, иногда более удобный способ задания больших групп символов. В языке RegEx в квадратных скобках могут встречаться специальные выражения, обозначающие сразу группу символов:

- [:alpha:] – буква;
- [:digit:] – цифра;
- [:alnum:] – буква или цифра;
- [:space:] – пробельный символ;
- [:punct:] – знак пунктуации.

Отрицательные группы

Иногда, когда альтернативных символов много, бывает довольно утомительно перечислять их все в квадратных скобках, особенно если подходят все символы, кроме нескольких. В этом случае следует воспользоваться конструкцией "[^]", которая обозначает любой символ, кроме тех, что перечислены после "[^" и до "]". Например, выражение "м[^ао]х" будет соответствовать всем строкам, содержащим буквы "м" и "х", разделенные любым символом, кроме "а" или "о".

Управляющие конструкции

Квантификаторы повторений

Перейдем к рассмотрению так называемых квантификаторов – спецсимволов, использующихся для уточнения действия предшествующих им символов первого класса.

Ноль и более совпадений. Звездочка "*" обозначает, что предыдущий символ может быть повторен ноль или более раз. Например, выражение "19*8" соответствует строке, в которой есть цифра "1", затем ноль или более цифр "9" и, наконец, "8".

Одно и более совпадений. Символ плюса "+" обозначает одно или более совпадений предшествующего символа или группы. Вот пример выражения, которое определяет слова, написанные через дефис: "[а-я]+-[а-я]+".

Ноль или одно совпадение. Иногда используют еще один квантификатор – знак вопроса "?". Он обозначает, что предыдущий символ может быть повторен ноль или один (но не более!) раз. Например, выражению "Петров[аы]?" будут соответствовать строки "Петров", "Петрова" и "Петровы".

Заданное число совпадений. Последний квантификатор повторения – фигурные

скобки "{}". С его помощью можно реализовать все перечисленные выше возможности. Существует несколько форматов его записи:

- A{n,m} – указывает, что символ "A" может быть повторен *от* n *до* m раз;
- A{n} – символ "A" должен быть повторен *ровно* n раз;
- A{n, } – символ "A" может быть повторен n *или более* раз.

Оператор альтернативы

При описании простых символов была рассмотрена конструкция "[…]", которая позволяла указывать, что в нужном месте строки должен стоять один из указанных символов. Это ни что иное, как оператор альтернативы, работающий с отдельными символами.

В языке RegEx есть возможность задавать альтернативы не одиночных символов, а сразу их групп. Это делается при помощи оператора "|". Вот несколько примеров его работы:

- "1|2|3" – полностью эквивалентно выражению [123];
- "^пре|^пере" – строки, которые начинаются с "пре" или "пере";
- "давать|давал|давала|давало|давали" – соответствует подстрокам, разделенным символом альтернативы "|".

Группирующие скобки

В примере "давать|давал|давала|давало|давали" подстрока "дава" встретилась в выражении пять раз. Для управления оператором альтернативы существуют группирующие круглые скобки "()". С их помощью выражение из последнего примера можно было записать так: "дава(ть|л|ла|ло|ли)". Скобки могут иметь произвольный уровень вложенности.

Мнимые символы

Мнимые символы – это просто участок строки между соседними символами, удовлетворяющий некоторым свойствам. Фактически, мнимый символ – это некая позиция в строке. Например, символ "^" соответствует началу строки, а "$" – ее концу.

Например, выражение "^пере" будет соответствовать любой строке, начинающейся на "пере", выражение "ть$" – строке оканчивающейся на "ть", а выражение "^перенять$" – точному совпадению со строкой "перенять".

Практическая часть (вопросы и задания для собеседования)

Примеры запросов

В приведенных ниже примерах наглядно продемонстрированы элементы языка запросов корпусного менеджера Bonito.

Задание 1. Поиск конкретной словоформы

В окно запроса вводится словоформа "run". Выдается:

announced that he would not <**run**> for reelection . Georgia
medical benefits paid out would <**run**> 1 billion or more in the
May , said today Jones will <**run**> well ahead of his GOP opponents
reports that he had decided to <**run**> and wanted Mr. Screvane ,
investigation Street car tracks <**run**> down the center of Pennsylvania

Система ищет полное соответствие запрашиваемому слову и выдает результат. Иных словоформ для конкретной словоформы "run" не будет найдено.

Задание 2. Поиск синтагмы

2.1. В окно запроса вводится "run in". Выдается:

contest . The Orioles got a <**run in**> the first inning when Breeding
record in the 600 - yard <**run in**> the Knights of Columbus track
The Bears added their last <**run in**> the sixth on Alusik 's double
for the third Indianapolis <**run in**> the ninth . Despite the 45
's first major league home <**run in**> the fifth put the Sox back

Поиск словоформ осуществляется в строго заданном (линейном) порядке, как неразрывная синтагма.

2.2. Допустим, нужно найти разрывную синтагму "take (smth) out".

В окно запроса вводится "take". Строится конкорданс для данной конкретной словоформы. Выбирается тип запроса Положительный фильтр (P-filter). В оба окна "From:" и "To:" вводится значение "2", что соответствует второй позиции справа от найденного слова для "оторванной" части синтагмы (в нашем примере "out"). Вокнозапросавводим "out". Выдается:

for governor would force it to <**take**> petitions **out** into voting
the peasant . Nonetheless , they <**take**> time **out** -- much time --
Mister McBride . You do that or <**take**> you **out** a permit right now

Разумеется, можно придумать и более сложные варианты подобных запросов с неоднократным применением Положительного фильтра.

Задание 3. Поиск различных форм слова

В окно запроса вводится "runs? in".

В данном запросе используется *управляющий символ* "?", который означает, что предшествующая ему буква "s" может встретиться ноль или один раз. Полученныйрезультатподтверждаетэто. Выдается:

tied the game , and single <**runs in**> the eighth and ninth gave
record in the 600 - yard <**run in**> the Knights of Columbus track
their eight hits for two <**runs in**> the sixth . Chuck Hinton
The Bears added their last <**run in**> the sixth on Alusik 's double
's first major league home <**run in**> the fifth put the Sox back

Задание 4. Поиск различных форм слова

В окно запроса вводится "run(|s|ning)".

Здесь используются группирующие скобки и оператор альтернативы (|) (логическое "или"). Системе дается команда найти конкретные словоформы "run" или "runs" или "running". Выдается:

announced that he would not **<run>** for reelection . Georgia
medical benefits paid out would **<run>** 1 billion or more in the
the group are interested in **<running>** on the required non -
lawyer and former FBI man is **<running>** against the Republican
tied the game , and single **<runs>** in the eighth and ninth gave

Задание 5. Поиск всех форм слова по лемме

В окно запроса вводится "[lemma="be"] within <head>". Выдается:

<head>DECISIONS **<ARE>** MADE</head>Asked to elaborate
<head>LEADERSHIP **<IS>** HOPEFUL</head>The housing
Nations .<head>FORMULA **<IS>** DUE THIS WEEK</head>The
year .<head>COULD **<BE>** SCRAMBLE</head>Some predict
ends .<head>CHOICE **<WAS>** EXPECTED</head>The selection
<head>TOBACCO ROAD **<IS>** DEAD . LONG LIVE TOBACCO

Так можно не только искать все словоформы по лемме, но и находить их в заданных полях документа (в данном примере в заголовочном поле, обозначенном тэгом <head>). Соответственно, если ввести несколько лемм подряд, то можно получить все варианты таких словосочетаний.

Задание 6. Поиск по морфологическим признакам

В окно запроса вводится "[tag="VVZv"]".

Выдается:

charge of the election , " **<deserves>** the praise and thanks of the
However , the jury said it **<believes>** " these two offices should be
of Fulton County , which **<receives>** none of this money " . The
when the new management **<takes>** charge Jan. 1 the airport be
face is a state law which **<says>** that before making a first

Пример демонстрирует замечательную возможность корпусного менеджера искать словоформы по морфологическим признакам. Код "VVZv" означает, что это третье лицо единственного числа (Zv) значимого глагола (VV). Такая кодировка предложена схемой аннотирования SUSANNE. Следовательно, данная возможность будет успешно использоваться, в первую очередь, теми, кто знаком с принципами данной схемы аннотирования.

Задание 7. Отображение морфологических признаков и леммы

В пункте командного меню выбирается "**View** **Attributes…**" и отмечаются пункты "**lemma**" и "**tag**".

В окно запроса вводится "[lemma="be"]".

Выдается:

in which the election <**was/be/VBDZ**> conducted . The September

-October term jury had <**been/be/VBN**> charged by Fulton Superior

stration and election laws " <**are/be/VBR**> outmoded or inadequate

" these two offices should <**be/be/VB0**> combined to achieve greater

Department, the jury said, " <**is/be/VBZ**> lacking in experienced

В конкордансе для каждого вхождения словоформы показана ее исходная форма и ряд морфологических признаков в виде кода.

Список рекомендуемой литературы

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk

7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. **Vi Improved - http://www.vim.org**

Практическое занятие 8.

Тема: **Регулярные выражения в автоматической обработке текста**

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

По материалам https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D1%8B%D0%B5_%D0%B2%D1%8B%D1%80%D0%B0%D0%B6%D0%B5%D0%BD%D0%B8%D1%8F

Регулярные выражения (англ. *regularexpressions*) — формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов (символов-джокеров, англ. *wildcardcharacters*). Для поиска используется строка-образец (англ. *pattern*, по-русски её часто называют «шаблоном», «маской»), состоящая из символов и метасимволов и задающая правило поиска. Для манипуляций с текстом дополнительно задаётся строка замены, которая также может содержать в себе специальные символы.

Регулярные выражения произвели прорыв в электронной обработке текстов в конце XX века. Набор утилит (включая редактор sed и фильтр grep), поставляемых в дистрибутивахUNIX, одним из первых способствовал популяризации регулярных

выражений для обработки текстов. Многие современные языки программирования имеют встроенную поддержку регулярных выражений. Среди них ActionScript, Perl, Java[1],PHP, JavaScript, языки платформы.NET Framework, Python, Tcl, Ruby, Lua, Gambas, C++ (стандарт 2011 года), Delphi, D и другие.

Регулярные выражения используются некоторыми текстовыми редакторами и утилитами для поиска и подстановки текста. Например, при помощи регулярных выражений можно задать шаблоны, позволяющие:

- найти все последовательности символов *«кот»* в любом контексте, как то: *«кот», «котлета», «терракотовый»*;
- найти отдельно стоящее слово *«кот»* и заменить его на *«кошка»*;
- найти слово *«кот»*, которому предшествует слово *«персидский»* или *«чеширский»*;
- убрать из текста все предложения, в которых упоминается слово *кот* или *кошка*.

Регулярные выражения позволяют задавать и гораздо более сложные шаблоны поиска или замены.

Результатом работы с регулярным выражением может быть:

- проверка наличия искомого образца в заданном тексте;
- определение подстроки текста, которая сопоставляется образцу;
- определение групп символов, соответствующих отдельным частям образца.

Если регулярное выражение используется для замены текста, то результатом работы будет новая текстовая строка, представляющая из себя исходный текст, из которого удалены найденные подстроки (сопоставленные образцу), а вместо них подставлены строки замены (возможно, модифицированные запомненными при разборе группами символов из исходного текста). Частным случаем модификации текста является удаление всех вхождений найденного образца — для чего строка замены указывается пустой.

Практическая часть (вопросы и задания для собеседования)

Разработайте регулярные выражения поиска:

Ч. Найти число

- натуральное
- целое
- десятичную дробь
- в экспоненциальной записи: 5.9736e24, 9.109382e-31

П. Пароли, логины, адреса

- простой логин: последовательность букв и цифр (от 4-х), нач. с буквы, регистр не различается
- сложный логин: последовательность букв, цифр и знаков `.-_` , нач. с буквы (от 8 до 16), регистр различается

- сложный пароль: последовательность букв, цифр и знаков **. - _ + = ? ! @ # $ %**…, содержащая минимум одну заглавную букву и одну цифру
- адрес электронной почты
- комментарий HTML <!-- Комментарий любой длины -->

Б. Библиографические ссылки

- Ссылка на работу двух соавторов Иванов и Петров (2001: 35); (Иванов и Петров 2001)

Список рекомендуемой литературы

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2025.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2025 г.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org

5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. **Vi Improved - http://www.vim.org**

Практическое занятие 9.

Тема: **Регулярные выражения для замены текста**

Цель: овладение студентами специфическими практическими навыками и умениями, необходимыми для проведения автоматизированного анализа корусов текстов с применением современных компьютерных технологий и программных продуктов.

Актуальность: актуальность изучения темы определяется существованием практической необходимости в подготовке специалистов в области применения программных продуктов автоматизированного анализа текста для массовой обработки электронных текстовых массивов.

Реализуемые компетенции:

Индекс:	**Формулировка:**
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Теоретическая часть

Представление символов в регулярных выражениях по их коду

В некоторых случаях предпочтительно представление символов по их коду.

Представление	Пояснение	Кодировка
`\0`*n*	*n* — восьмеричное число от 0 до 377	8-битная
`\x`*dd*	*d* — шестнадцатеричная цифра	
`\u`*dddd*		16-битная (Юникод)

Управляющие символы

Представление	Символ	Обозначение	Расшифровка
`\t`	Табуляция	`HT`	Horizontal tabulation
`\v`	Вертикальная табуляция	`VT`	Vertical tabulation
`\r`	Возврат каретки	`CR`	Carriage return
`\n`	Перевод строки	`LF`	Line feed
`\f`	Конец страницы	`FF`	Form feed
`\a`	Звонок	`BEL`	Bell character
`\e`	Escape-символ	`ESC`	Escape character

\b	Забой	BS	Backspace
	Должен находиться внутри квадратных скобок (иначе интерпретируется как граница слова).		
\cA … \cZ	Ctrl+A … Ctrl+Z		
	Например, последовательность \cM\cJ соответствует управляющим символам CRLF.		
	Эквивалентно \x01 … \x1A.		

Сокращённые обозначения символьных классов

Для часто используемых символьных классов существуют краткие обозначения.

Представление	Эквивалент	Значение
\d	[0-9]	Цифра
\D	[^\d]	Любой символ, кроме цифры
\w	[A-Za-zA-Яа-я0-9_]	Символы, образующие «слово» (буквы, цифры и символ подчёркивания)[1]
\W	[^\w]	Символы, не образующие «слово»
\s	[\t\v\r\n\f]	Пробельный символ
\S	[^\s]	Не пробельный символ

Символьные классы POSIX

Многие диапазоны символов зависят от выбранных настроек локализации. POSIX стандартизовал объявление некоторых классов и категорий символов, как показано в следующей таблице.

POSIX-класс	Эквивалент	Значение
[:upper:]	[A-Z]	Символы верхнего регистра
[:lower:]	[a-z]	Символы нижнего регистра
[:alpha:]	[[:upper:][:lower:]]	Буквы
[:digit:]	[0-9], т. е. \d	Цифры
[:xdigit:]	[[:digit:]A-Fa-f]	Шестнадцатеричные цифры
[:alnum:]	[[:alpha:][:digit:]]	Буквы и цифры
[:word:]	[[:alnum:]_], т. е. \w	Символы, образующие «слово»
[:punct:]	[-!"#$%&'()*+,./:;<=>?@[\\\]_`{|}~]	Знаки пунктуации
[:blank:]	[\t]	Пробел и табуляция

`[:space:]`	`[[:blank:]\v\r\n\f]`, т. е. `\s`	Пробельные символы
`[:cntrl:]`	`[\x00-\x1F\x7F]`	Управляющие символы
`[:graph:]`	`[\x21-\x7E]`	Печатные символы
`[:print:]`	`[\x20-\x7E]`, т. е. `[[:graph:] ]`	Печатные символы с пробелом

Использование класса возможно лишь внутри квадратных скобок (пример частой ошибки — `^[:upper:]il+$` вместо `^[[:upper:]]il+$`).

Практическая часть (вопросы и задания для собеседования)

Разработайте регулярные выражения для поиска и замены:

Д. Дефисы, тире и т. п.

- дефис между двумя числами — на короткое тире
- дефис между двумя пробелами — на длинное тире
- дефис между знаком препинания и пробелом — на длинное тире
- прямые кавычки на «елочки»

С. Список временных меток для синхронизации звука с текстом

Исходныйформат:
Regions List: archi06-txo-01-KVS-0079.wav

Name	Start	End	Length
000	1.811156		
001	7.291066		
002	15.464490		
003	23.173515		
004	31.764898		
005	38.034286		

...

Требуемыйформат:

```
<synch-list>
<span class="synch">1.811</span><!--  000 -->
<span class="synch">7.291</span><!--  001 -->
<span class="synch">15.464</span><!--  002 -->
<span class="synch">23.173</span><!--  003 -->
<span class="synch">31.764</span><!--  004 -->
<span class="synch">38.034</span><!--  005 -->
...
</synch-list>
```

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. -

URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

7. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
8. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
9. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
10. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
11. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
12. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Методическая литература:

1. Каменский М.В. Методические рекомендации по организации самостоятельной работы студентов по дисциплине «Корпусная лингвистика».– Ставрополь, 2026.
2. Каменский М.В. Методические указания по выполнению практических работ по дисциплине «Корпусная лингвистика». – Ставрополь, 2026.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. ViImproved - http://www.vim.org

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Методические указания

по выполнению самостоятельной работы студентов по дисциплине «Корпусная лингвистика»

Направление подготовки	45.04.02 Лингвистика
Направленность (профиль)	Современные методы прикладной лингвистики и перевода
Год начала подготовки	2026
Форма обучения	Очная
Реализуется в семестре	1

Ставрополь
2026

Содержание

1. Введение

Методические рекомендации к самостоятельной работе студентов по дисциплине «Корпусная лингвистика» разработаны в соответствии с рабочей программой дисциплины по направлению 45.04.02 - Лингвистика, программа – Современные методы прикладной лингвистики и перевода.

Основной формой работы студента является не только работа на лекции, изучение конспекта лекций, их дополнение рекомендованной литературой, но и большая самостоятельная учебная работа, которая позволит глубоко проникнуть в суть рассматриваемой проблемы и подготовить почву для написания кандидатской диссертации. Но для успешной учебной деятельности, ее интенсификации необходимо учитывать следующие субъективные факторы:

1. Знание программного материала, наличие прочной системы знаний, необходимой для усвоения основных дисциплин, предусмотренных программой, общая совокупность которых обусловливает уровень овладения грамматическим компонентом иноязычной речи.
2. Наличие выработанных умений, навыков умственного труда:

а) умение делать глубокий, обстоятельный анализ при работе с книгой, Интернет–источниками;

б) владение логическими операциями: сравнение, анализ, обобщение, определение понятий, правила систематизации и классификации.

3. Специфика познавательных психических процессов: внимание, память, речь, наблюдательность, интеллект и мышление.
4. Хорошая работоспособность, которая обеспечивается нормальным физическим состоянием.
5. Соответствие избранной деятельности, профессии индивидуальным способностям. Необходимо выработать умение саморегулировать свое эмоциональное состояние и устранять обстоятельства, нарушающие деловой настрой, мешающие намеченной работе.
6. Овладение оптимальным стилем работы, обеспечивающим успех в деятельности.
7. Уровень требований к себе, определяемый сложившейся самооценкой.

Адекватная оценка знаний, достоинств, недостатков – важная составляющая самоорганизации человека, без нее невозможна успешная работа по управлению своим поведением, деятельностью.

По наблюдениям исследователей педагогов, одна из основных особенностей обучения заключается в том, что постоянный внешний контроль заменяется самоконтролем, активная роль в обучении принадлежит уже не столько преподавателю, сколько студенту.

2. Общая характеристика самостоятельной работы студента при изучении дисциплины

Самостоятельная работа студента в рамках дисциплины «Корпусная лингвистика» понимается как планируемая учебная работа, выполняемая во внеаудиторное (аудиторное) время по заданию и при методическом руководстве преподавателя, но без его непосредственного участия.

Самостоятельная работа направлена на формирование следующих компетенций:

Реализуемые компетенции:

Индекс:	Формулировка:
ОПК-6	Способен применять современные технологии при осуществлении сбора, обработки и интерпретации данных эмпирического исследования; составлять и оформлять научную документацию.
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

Цель дисциплины– научить студента осмысленно и самостоятельно работать: 1) с учебным материалом по дисциплине, 2) с научной информацией, актуальными исследованиями в области лингвистики, 3) с эмпирическими данными, получаемыми в ходе экспериментальных лингвистических исследований, 4) с методологическими подходами современных лингвистических исследований; 5) с конкретными лингвистическими методами и методиками.

Задачи самостоятельной работы:

- систематизировать и закрепить полученные теоретические знания и практические умения студентов;
- развить познавательные способности и активность студентов: творческую инициативу, самостоятельность, ответственность и организованность;
- сформировать и развить навыки ведения самостоятельной работы и овладения методикой исследования при решении разрабатываемых в учебной деятельности проблем и вопросов;
-повысить уровень подготовленности к самостоятельной работе в соответствии с выбранным научным направлением в условиях современного состояния науки и культуры.

Таким образом, самостоятельная работа приобщает научному и исследовательскому творчеству, поиску и анализу актуальных проблем современной лингвистической науки.

3. План-график выполнения самостоятельной работы.

Коды реализуемых компетенций	Вид деятельности студентов	Итоговый продукт самостоя-тельной работы	Средства и технологии оценки	Объем часов, в том числе (астр.)		
				СРС	Контактная работа с преподавателем	Всего
1 семестр						
ОПК-6, УК-1	Подготовка к практическим занятиям	Конспект	Собеседование	10	1,5	11,5
ОПК-6, УК-1	Подготовки к лабораторным занятиям	Конспект	Собеседование	10	0,5	10,5
ОПК-6, УК-1	Самостоятельное изучение литературы	Доклад, презентация	Собеседование	15	2	17
ОПК-6, УК-	Написание	Статья,	Собеседова	15	2	17

1	статьи, тезисов	тезисы	ние			
ОПК-6, УК-1	Выполнение исследовательского проекта по заданной проблематике	Проект	Собеседование	30,6	3,4	34
Итого за 1 семестр				81	9	90
Итого				81	9	90

Для выполнения самостоятельной работы необходимо пользоваться литературой, которая предложена в списке рекомендуемой литературы, Интернет-ресурсами или другими источниками по усмотрению студента.
Самостоятельная работа рассчитана на разные уровни мыслительной деятельности. Выполненная работа позволит приобрести не только знания, но и умения, навыки, а также выработать свою методику подготовки, что очень важно в дальнейшем процессе научной деятельности.
При изучении дисциплины предусматриваются следующие формы самостоятельной работы студента:

2) самостоятельное изучение основной и дополнительной литературы по дисциплине с конспектированием по разделам;
3) работа с электронными ресурсами в сети Интернет;
4) конспектирование и реферирование первоисточника и научно-исследовательской литературы;
5) подготовка к семинару-круглому столу;
6) подготовка мультимедийной презентации;
7) подготовка доклада.

4. Контрольные точки и виды отчетности по ним
Не предусмотрены

5. Методические рекомендации по изучению теоретического материала
Чтение основной и дополнительной литературы по курсу с конспектированием по разделам.
Самостоятельная работа при чтении учебной литературы начинается с изучения конспекта материала, полученного при слушании лекций преподавателя. Полученную информацию необходимо осмыслить. При необходимости, в конспект лекций могут быть внесены схемы, другая дополнительная информация. При изучении нового материала составляется конспект. Сжато излагается самое существенное в данном материале.

Работа с электронными ресурсами в сети Интернет.
Для повышения эффективности самостоятельной работы студент должен уметь работать в поисковой системе сети Интернет и использовать найденную информацию при подготовке к занятиям. Поиск информации можно вести по автору, заглавию, виду издания, году издания или издательству. Также в сети Интернет доступна услуга по скачиванию методических указаний и учебных пособий, подбору необходимой научной литературы.

Конспектирование и реферирование первоисточника и научно-исследовательской литературы.

Конспект представляет собой дословные выписки из текста источника. При этом необходимо понимать, что конспект – это не полное переписывание чужого текста. Необходимо знать, что при написании конспекта сначала прочитывается текст – источник, в нём выделяются основные положения, подбираются примеры, идёт перекомпоновка материала, а уже затем оформляется текст конспекта. Конспект может быть полным, когда работа идёт со всем текстом источника или неполным, когда интерес представляет какой-либо один или несколько вопросов, затронутых в источнике.

Реферирование — это сложный творческий процесс, в основе которого лежит умение выделить главную информацию из текста первоисточника. Реферирование – процесс аналитически-синтетического обработки информации, которая заключается в анализе первичного документа, нахождении значимых в смысловом отношении данных (основных положений, фактов, доведите день, результатов, выводов) Реферирование имеет целью сократить физический объем первичного документа при сохранении его основного смыслового содержания, используется в научной, издательской, информационной и библиографической деятельности.

6. Методические указания (по видам работ, предусмотренных рабочей программой дисциплины)

Подготовка к круглому столу

Подготовка к семенару-круглому столу начинается с распределение форм участия и функции студентов в семинаре-круглом столе. Студентами осуществляется определение круга проблем и вопросов, подлежащих обсуждению; подбор основной и дополнительной литературы к теме семинара - круглого стола, а также дальнейшее изучение литературы.

Подготовка мультимедийной презентации

Презентация, согласно толковому словарю русского языка Д.Н. Ушакова: «... способ подачи информации, в котором присутствуют рисунки, фотографии, анимация и звук». Для подготовки презентации рекомендуется использовать LibreOffice Impress (для подготовки собственно мультимедийных презентаций) и LibreOffice Writer (для составления текстового сопровождения презентации), являющихся компонентами открытого и свободного офисного пакета LibreOffice. Также допускается использование проприетарного продукта Microsoft Office (Powerpoint и Word, соответственно), однако в этом случае должны использоваться наиболее совместимые форматы .ppt, .doc (но не .pptx, .docx).

Для подготовки презентации необходимо собрать и обработать начальную информацию.

Последовательность подготовки презентации:

1. Четко сформулировать цель презентации: вы хотите свою аудиторию мотивировать, убедить, заразить какой-то идеей или просто формально отчитаться.
2. Определить каков будет формат презентации: живое выступление (тогда, сколько будет его продолжительность) или электронная рассылка (каков будет контекст презентации).
3. Отобрать всю содержательную часть для презентации и выстроить логическую цепочку представления.
4. Определить ключевые моменты в содержании текста и выделить их.
5. Определить виды визуализации (картинки) для отображения их на слайдах в соответствии с логикой, целью и спецификой материала.
6. Подобрать дизайн и форматировать слайды (количество картинок и текста, их расположение, цвет и размер).
7. Проверить визуальное восприятие презентации.

К видам визуализации относятся иллюстрации, образы, диаграммы, таблицы.
Иллюстрация – представление реально существующего зрительного ряда.
Образы – в отличие от иллюстраций – метафора. Их назначение –вызвать эмоцию и создать отношение к ней, воздействовать на аудиторию. С помощью хорошо продуманных и представляемых образов, информация может надолго остаться в памяти человека.
Диаграмма – визуализация количественных и качественных связей. Их используют для убедительной демонстрации данных, для пространственного мышления в дополнение к логическому.
Таблица – конкретный, наглядный и точный показ данных. Ее основное назначение – структурировать информацию, что порой облегчает восприятие данных аудиторией.

Практические советы по подготовке презентации.

- готовьте отдельно: печатный текст + слайды + раздаточный материал;
- слайды – визуальная подача информации, которая должна содержать
- минимум текста, максимум изображений, несущих смысловую нагрузку, выглядеть наглядно и просто;
- текстовое содержание презентации – устная речь или чтение, которая
- должна включать аргументы, факты, доказательства и эмоции;
- рекомендуемое число слайдов 10-12;
- обязательная информация для презентации: тема, фамилия и инициалы
- выступающего; план сообщения; краткие выводы из всего сказанного; список использованных источников;
- раздаточный материал – должен обеспечивать ту же глубину и охват, что и живое выступление: люди больше доверяют тому, что они могут унести с собой, чем исчезающим изображениям, слова и слайды забываются, а раздаточный материал остается постоянным осязаемым напоминанием; раздаточный материал важно раздавать в конце презентации; раздаточный материалы должны отличаться от слайдов, должны быть более информативными.

Доклад, согласно толковому словарю русского языка Д.Н. Ушакова:
«... сообщение по заданной теме, с целью внести знания из дополнительной литературы, систематизировать материл, проиллюстрировать примерами, развивать навыки самостоятельной работы с научной литературой, познавательный интерес к научному познанию».
Тема доклада должна быть согласованна с преподавателем и соответствовать теме учебного занятия. Материалы при его подготовке, должны соответствовать научно-методическим требованиям вуза и быть указаны в докладе. Необходимо соблюдать регламент, оговоренный при получении задания. Иллюстрации должны быть достаточными, но не чрезмерными.
Работа студента над докладом-презентацией включает отработку умения самостоятельно обобщать материал и делать выводы в заключении, умения ориентироваться в материале и отвечать на дополнительные вопросы слушателей, отработку навыков ораторства, умения проводить диспут.
Докладчики должны знать и уметь: сообщать новую информацию; использовать технические средства; хорошо ориентироваться в теме всего семинарского занятия; дискутировать и быстро отвечать на заданные вопросы; четко выполнять установленный

регламент (не более 10 минут); иметь представление о композиционной структуре доклада и др.

Структура выступления

Вступление помогает обеспечить успех выступления по любой тематике. Вступление должно содержать: название, сообщение основной идеи, современную оценку предмета изложения, краткое перечисление рассматриваемых вопросов, живую интересную форму изложения, акцентирование внимания на важных моментах, оригинальность подхода.

Основная часть, в которой выступающий должен глубоко раскрыть суть затронутой темы, обычно строится по принципу отчета. Задача основной части – представить достаточно данных для того, чтобы слушатели заинтересовались темой и захотели ознакомиться с материалами. При этом логическая структура теоретического блока не должны даваться без наглядных пособий, аудиовизуальных и визуальных материалов.

Заключение – ясное, четкое обобщение и краткие выводы, которых всегда ждут слушатели

Написание доклада

Доклад – публичное сообщение, представляющее собой развёрнутое изложение определённой темы.

Этапы подготовки доклада:

1. Определение цели доклада.
2. Подбор необходимого материала, определяющего содержание доклада.
3. Составление плана доклада, распределение собранного материала в необходимой логической последовательности.
4. Общее знакомство с литературой и выделение среди источников главного.
5. Уточнение плана, отбор материала к каждому пункту плана.
6. Композиционное оформление доклада.
7. Заучивание, запоминание текста доклада, подготовки тезисов выступления.
8. Выступление с докладом.
9. Обсуждение доклада.
10. Оценивание доклада

Композиционное оформление доклада – это его реальная речевая внешняя структура, в ней отражается соотношение частей выступления по их цели, стилистическим особенностям, по объёму, сочетанию рациональных и эмоциональных моментов, как правило, элементами композиции доклада являются: вступление, определение предмета выступления, изложение (опровержение), заключение.

Вступление помогает обеспечить успех выступления по любой тематике.

Вступление должно содержать:

1. название доклада;;
2. сообщение основной идеи;
3. современную оценку предмета изложения;
4. краткое перечисление рассматриваемых вопросов;
5. интересную для слушателей форму изложения;
6. акцентирование оригинальности подхода.

Выступление состоит из следующих частей:

Основная часть, в которой выступающий должен раскрыть суть темы, обычно строится по принципу отчёта.

Задача основной части: представить достаточно данных для того, чтобы слушатели заинтересовались темой и захотели ознакомиться с материалами.

Заключение - это чёткое обобщение и краткие выводы по излагаемой теме.

7. Список литературы, использованной при составлении методических рекомендаций

Основная литература:

1. Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463

Дополнительная литература:

1. 5th GATE Training Course. - The University of Sheffield, 2012. - http://gate.ac.uk/wiki/TrainingCourseJune2012/
2. Захаров, В.П. Корпусная лингвистика: Учебник для студентов гуманитарных вузов : учебник / В.П. Захаров, С.Ю. Богданова. - Иркутск : Иркутский государственный лингвистический университет, 2014. - 161 с. - ISBN 978-5-88267-316-0 ; То же [Электронный ресурс]. - URL:http://biblioclub.ru/index.php?page=book&id=89753
3. Каменский М.В. Когнитивно-функциональная модель дискурсных маркеров. – Ставрополь: Изд-во СКФУ, 2014. – 186 с.
4. Роббинс, А. Изучаем редакторы vi и Vim / А. Роббинс, Э. Хана, Л. Лэмб. – М.: Символ-Плюс, 2013. – 512 с.
5. Friedl, J.E.F. Mastering Regular Expressions / J.E.F. Friedl. – 3rd. ed. O'Reilly Media, 2012. – 544 p.
6. Horstmann, C., Cornell, G. Core Java Volume I – Fundamentals (9th Edition) / C. Horstmann, G. Cornell. – Prentice Hall, 2012. – 1008 p.

Интернет-ресурсы:

1. Applied Linguistics – http://www.appliedlinguistics.org
2. Center for Applied Linguistics – http://www.cal.org
3. GATE (Sheffield University, UK) – http://gate.ac.uk

Программное обеспечение:

1. Annotation Graph Toolkit (AGTK) – http://agtk.sourceforge.net
2. Applied Linguistics – http://www.appliedlinguistics.org
3. Center for Applied Linguistics – http://www.cal.org
4. Code::Blocks C/C++ IDE – http://www.codeblocks.org
5. Emacs – http://www.gnu.org/software/emacs
6. GATE (General Architecture for Text Engineering) – http://gate.ac.uk
7. GNU Compiler Collection – http://gcc.gnu.org
8. Helsinki Finite-State Technology – http://sourceforge.net/projects/hfst
9. Leopar Language Parser – http://leopar.loria.fr
10. NLTK (Natural Language Toolkit) – http://www.nltk.org
11. Open Natural Language Processing (OpenNLP) – http://opennlp.sourceforge.net
12. Praat – http://www.fon.hum.uva.nl/praat
13. Python Programming Language – http://www.python.org
14. Sed – http://www.gnu.org/s/sed
15. Sonic Visualizer – http://www.sonicvisualiser.org
16. Vi Improved - http://www.vim.org