**Министерство образования Республики Беларусь**

**Учебно-методическое объединение по гуманитарному образованию**

**УТВЕРЖДАЮ**

Первый заместитель Министра образования

Республики Беларусь

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А. И. Жук

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2012 г.

Регистрационный № ТД-\_\_\_\_\_\_\_\_/ тип.

**КОРПУСНАЯ ЛИНГВИСТИКА**

Типовая учебная программа

для высших учебных заведений по специальностям

1-21 05 01 «Белорусская филология» (по направлениям)

Направление специальности 1-21 05 01-02 «Белорусская филология (компьютерное обеспечение)»

1-21 05 02 «Русская филология» (по направлениям)

Направление специальности 1-21 05 02-02 «Русская филология   
(компьютерное обеспечение)»

СОГЛАСОВАНО СОГЛАСОВАНО

Председатель Начальник Управления

Учебно-методического объединения высшего и среднего специального образования

по гуманитарному образованию Министерства образования Республики Беларусь

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В. Л. Клюня \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ С. И. Романюк

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2012 г. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2012 г.

Проректор по научно-методической работе

Государственного учреждения образования

«Республиканский институт высшей школы»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ И. В. Титович

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2012 г.

Эксперт-нормоконтролер

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2012 г.

Минск 2012

**СОСТАВИТЕЛИ:**

**Рычкова Людмила Васильевна** – заведующий кафедрой общего и славянского языкознания учреждения образования «Гродненский государственный университет имени Янки Купалы», кандидат филологических наук, доцент

**Станкевич Алеся Юрьевна** – старший преподаватель кафедры общего и славянского языкознания учреждения образования «Гродненский государственный университет имени Янки Купалы»

**РЕЦЕНЗЕНТЫ:**

кафедра интеллектуальных информационных технологий факультета информационных технологий и управления учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

**Коваль Владимир Иванович** – заведующий кафедрой русского, общего и славянского языкознания учреждения образования «Гомельский государственный университет имени Франциска Скорины», доктор филологических наук, профессор

**РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ В КАЧЕСТВЕ ТИПОВОЙ:**

кафедрой общего и славянского языкознания филологического факультета учреждения образования «Гродненский государственный университет имени Янки Купалы» (протокол № 5 от 11 мая 2012 г.);

Научно-методическим советом Белорусского государственного университета (протокол № 5 от 28 мая 2012 г.);

Научно-методическим советом по филологическим специальностям учебно-методического объединения по гуманитарному образованию (протокол № 6 от 21 июня 2012 г.).

**Ответственный за редакцию**: Л. В. Рычкова

**Ответственный за выпуск**: Л. В. Рычкова

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

**1.1. Характеристика учебной дисциплины**

Дисциплина «Корпусная лингвистика» входит в состав обязательных дисциплин, предусмотренных образовательными стандартами филологических специальностей в рамках направления «Компьютерное обеспечение».

Содержательное ядро дисциплины составляет проблематика такой интердисциплинарной отрасли прикладных лингвистических знаний, как корпусная лингвистика. Предполагается, что студенты к моменту изучения данной дисциплины уверенно овладели навыками пользования типовым программным и информационным обеспечением, полученными ими при изучении дисциплин «Основы информатики» и «Введение в компьютерную филологию», а также овладели основным понятийным аппаратом как компьютерной лингвистики, так и «зонтичной» области компьютерной филологии. Овладение дисциплиной необходимо для непосредственной поддержки профессиональной деятельности специалиста по компьютерной филологии.

**1.2. Цель и задачи учебной дисциплины**

**Цель дисциплины**  – сформировать у студентов системное представление о методах и технологиях корпусной лингвистики, а также о направлениях применения методов и инструментария корпусной лингвистики для решения профессиональных задач в области филологии.

Основные **задачи** дисциплины включают:

– выработку представлений о типологии лингвистически корпусов, технологиях их создания и об областях применения различных видов корпусов (теоретическая составляющая);

– выработку практических навыков решения задач по разделам курса (собственно прикладная, или практико-ориентированная составляющая).

В результате изучения учебной дисциплины студент должен:

**знать**:

– терминологический аппарат изучаемой предметной области;

– классы лингвистических задач, решаемых с применением корпусных технологий;

– основные этапы создания лингвистических корпусов;

– характерные особенности доступных on-line корпусов языков специальности;

**уметь**:

– решать профессиональные задачи с обращением к методам и инструментам корпусной лингвистики;

– работать с изучаемыми в пределах курса инструментами корпусной лингвистики;

– работать с доступными on-line корпусами различных языков;

–  применять лингвистические корпусы как исследовательский и дидактический инструмент.

Основные **компетенции**, которые должны быть сформированы у студентов:

*– компетенция в сфере познавательной деятельности:* применение методов корпусной лингвистики для решения профессионально-ориентированных задач;

*– профессиональные компетенции:* владение корпусными технологиями для решения задач извлечения, обработки и анализа информации; владение приемами работы с программными средствами корпусной лингвистики.

**1.3. Рекомендуемые формы и методы обучения и воспитания** включают:

- интерактивные лекции с использованием современных технических средств наглядности и разнообразных форм дидактического диалога;

- практические и лабораторные занятия с включением подсистемы контролируемых самостоятельных заданий.

Основной метод обучения: наглядно-практический.

**1.4. Рекомендации по организации самостоятельной работы студентов**

Формой проведения самостоятельной работы студентов, оптимальной при изучении данной дисциплины, является выполнение контрольно-обучающих заданий средствами заданного программного инструмента, разработка собственных проектов (мини-корпусов или их компонентов) либо участие в разработке некоммерческого открытого корпусного проекта.

Рекомендуемая форма итогового контроля знаний и умений студентов - экзамен.

**1.5. Рекомендуемый объем часов и формы проведения занятий**

На изучение дисциплины «Корпусная лингвистика» отводится 80 часов, из них аудиторных – 34 часов.

Примерное распределение аудиторного времени следующее:

– лекции – 20 часов;

– практические занятия – 14 часов.

**ПРИМЕРНЫЙ ТЕМАТИЧЕСКИЙ ПЛАН**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№**  **п/п** | **Название раздела**  (темы, занятия) | **Количество аудиторных часов** | | | | |
| **лекции** | **практические занятия** | **лабораторные занятия** | **семинары** | **управляемая (контролируемая) самостоятельная работа студента** |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 1. | Корпус как особый лингвистический ресурс | 2 |  |  |  |  |
| 2. | Корпусная лингвистика в динамическом аспекте | 2 |  |  |  | 2 |
| 3. | Типология корпусов и направления их использования | 4 |  |  |  | 4 |
| 4. | Web как корпус | 2 | 2 |  |  | 4 |
| 5. | Лингвистические задачи, решаемые с применением корпусных технологий | 2 | 2 |  |  | 2 |
| 6 | Корпусы языка специальности, доступные on-line | 2 | 2 |  |  | 16 |
| 7. | Этапы создания корпуса.  Разметка. Средства разметки | 2 | 2 |  |  | 2 |
| 8. | Корпус-менеджер как инструментальное средство ведения корпуса | 2 | 4 |  |  | 8 |
| 9. | Международные стандарты, применяемые в корпусной лингвистике | 2 | 2 |  |  | 8 |
|  | **ИТОГО ЧАСОВ ПО КУРСУ:** | **20** | **14** |  |  | **46** |

**СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА**

**Раздел 1. Основы корпусной лингвистики**

**Корпус как особый лингвистический ресурс.** Определение лингвистического корпуса. Понятия размеченного корпуса, корпуса первого порядка («сырого» корпуса). Проблема репрезентативности и полноты корпуса. Структура корпуса: корпус данных, корпус-менеджер.

**Корпусная лингвистика в динамическом аспекте.** Корпусы «первого поколения»: Брауновский корпус | The Brown Corpus, Корпус LOB | The Lancaster-Oslo-Bergen Corpus. Мегакорпусы, их особенности: Британский национальный корпус | British National Corpus | BNC; Чешский национальный корпус, Национальный корпус русского языка | НКРЯ.

**Типология корпусов и направления их использования.** Типы лингвистических корпусов и основания их выделения. Специальные виды корпусов. Параллельный корпус: особенности разработки (методы выравнивания), области применения. Сопоставительные корпусы, их отличие от параллельных. Корпус текстов обучающихся (Learner corpus): особенности разработки (разметка ошибок), области применения. Исторические корпусы: особенности разработки (представление материала), области применения.

**Web как корпус**. Сервисы типа Web as Corp vs Интернет-ИПС. Сервисы типа Web as Corp: сервис Web Corp, сервис Querying Internet corpora С.А. Шарова.

**Лингвистические задачи, решаемые с применением корпусных технологий.** Корпусы как средство получения статистик, словоуказателей, конкордансов. Корпусные технологии в компьютерной лексикографии. Корпусные технологии в обучении иностранным языкам. Использование корпусов в системах АОТ.

**Корпусы языка специальности, доступные on-line:** виды разметки, выходные интерфейсы и возможности корпус-менеджеров.

Открытые корпусные проекты (на примере проекта OpenCorpora, Корпус текстов петербургских школьников | SPb EFL Learner Corpus).

**Раздел. 2. Технологии корпусной лингвистики**

**Этапы создания корпуса. Разметка.** Разметка как научная деятельность. Виды разметки: экстралингвистическая разметка|метаразметка, структурная разметка, лингвистическая разметка Виды лингвистической разметки: морфологическая, синтаксическая, семантическая, анафорическая, просодическая. Разметка ручная и автоматическая; их преимущества и недостатки. **Средства разметки.** Языковые средства разметки: SGML, XML. Инструментальные средства разметки: SGML- и XML-редакторы, конструкторы схем (на примере TEI Roma), средства автоматической морфологической и синтаксической разметки.

**Корпус-менеджер как инструментальное средство ведения корпуса:** основные функции корпус-менеджера; применимость корпус-менеджеров для обработки корпусов первого порядка и размеченных корпусов (на примерах AntConc , SARA (XAIRA), DDC).

**Международные стандарты, применяемые в корпусной лингвистике.** Объекты стандартизации. Международные стандарты для представления корпусов текстов (проект TEI | Text Encoding Initiative, рекомендации EAGLES | Expert Advisory Group on Language Engineering Standards).

TEI LIGHT как подмножество системы кодирования текстов TEI: назначение, принципы отбора элементов подмножества. Подмножества TEI, доступные через интерфейс конструктора TEI ROMA (на примере **TEI for Linguistic Corpora**).

**ИНФОРМАЦИОННО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ  
ПО ДИСЦИПЛИНЕ**

**4.1.1. Перечень рекомендуемой литературы**

1. Баранов, А.Н. Введение в прикладную лингвистику / А.Н. Баранов. – М., 2001. – 360 с.
2. Беляева, Л.Н. Лексикографический потенциал параллельного корпуса текстов /Л.Н. Беляева // Труды международной конференции «Корпусная лингвистика – 2004». – СПб., 2004. – С. 55 – 64.
3. Добрушина, Е.Н. Использование НКРЯ в преподавании филологических дисциплин: Методические разработки в помощь преподавателям высшей и средней школы. – М., 2008.
4. Захаров, В.П. Корпусная лингвистика: учебник для студентов гуманитарных вузов / В.П. Захаров, С.Ю. Богданова. – Иркутск: ИГЛУ, 2011. – 161 с.
5. Захаров, В.П. Веб-пространство как языковой корпус / В.П. Захаров // Труды международной конференции «Диалог-2005». – М., 2005. – С. 166–171 [http://www.dialog-21.ru/Archive/2005/Zakharov%20V/ZakharovV.pdf].
6. Зубов, А.В. Информационные технологии в лингвистике: учеб. пособие /А.В. Зубов, И.И. Зубова.– М. : Академия, 2004. – 208 с.
7. Камшилова, О.Н. Исследовательский потенциал корпуса английских текстов петербургских школьников: анализ интерязыка / О.Н. Камшилова // Известия Российского государственного педагогического университета имени А.И.Герцена [Текст]. – СПб., 2009. – N 104. – С.114 – 123 [ftp://lib.herzen.spb.ru/text/kamshilova\_114\_123.pdf].
8. Копотев, М.В. Современная корпусная русистика /М.В. Копотев, А. Мустайоки // Инструментарий русистики: корпусные подходы. – Хельсинки, 2008. – С. 7 – 24.
9. Национальный корпус русского языка: 2003–2005: результаты и перспективы: сб. ст. / Ин-т рус. яз. им. В.В. Виноградова Рос. Акад. наук. – М.: Индрик, 2005. – 343 с.
10. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / Отв. ред. В.А. Плунгян. – СПб.: Нестор-История, 2009.– 502 с.
11. НТИ. – Сер.2. Информационные процессы и системы.– 2003.– № 10. Корпусная лингвистика в России [http://www.viniti.ru/cgi-bin/nti/nti.pl?action=show&year=2\_2003&issue=10&%20%20%20page=2].
12. НТИ. – Сер.2. Информационные процессы и системы.– 2003.– № 6. Корпусная лингвистика в России [http://www.viniti.ru/cgi-bin/nti/nti.pl?action=show&year=2\_2003&issue=6&%20%20%20page=2].
13. Резникова, Т.И. Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) / Т.И. Резникова, М.В. Копотев // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. — М., 2005. – С. 31–61 [http://ruscorpora.ru/sbornik2005/04reznikova.pdf].
14. Biber, D. Corpus Linguistics. Investigating Language Structure and Use / D. Biber, S. Conrad, R. Reppen – Cambridge University Press, 1998.– 300 p.
15. Kennedy, G. An Introduction to Corpus Linguistics. – Longman, 1998.– 316 р.

**4.1.2. Дистрибутивы программных продуктов и on-line ресурсы**

1. Список корпусов на сайте Национального корпуса русского языка. – Режим доступа: http://www.ruscorpora.ru/corpora-other.html.– Дата доступа: 25.03.2012.
2. Learner corpora around the world [Electronic resource]. – Mode of access: http://www.uclouvain.be/en-cecl-lcworld.html.– Date of access: 27.11.2011.
3. Querying Internet corpora [Electronic resource] / S. Sharoff. – University of Leeds. – Mode of access: http://corpus1.leeds.ac.uk/internet.html.– Date of access: 25.03.2012.
4. WebCorp [Electronic resource] / Research and Development Unit for English Studies, Birmingham City University.– Mode of access: <http://www.webcorp.org.uk>/.– Date of access: 25.03.2012.
5. Корпус текстов петербургских школьников. SPb EFL Learner Corpus / Кафедра прикладной лингвистики РГПУ им. А. И. Герцена. – Режим доступа: http://www.spbeflcorp.ru/. – Дата доступа: 25.03.2012.
6. Проект «Открытый корпус» (OpenCorpora). – Режим доступа: http://opencorpora.org/.– Дата доступа: 25.03.2012.
7. AntConc [Electronic resource] / L. Antony. – Mode of access: http://www.antlab.sci.waseda.ac.jp/antconc\_index.html.– Date of access: 25.03.2012.
8. DDC-concordance: a search engine for linguists [Electronic resource].– Mode of access: http://www.ddc-concordance.org/. – Date of access: 25.03.2012.
9. Manatee / Bonito [Electronic resource]. – Mode of access: http://www.textforge.cz/download.– Date of access: 25.03.2012.
10. XAIRA: XML Aware Indexing and Retrieval Architecture [Electronic resource] / University of Oxford.– Mode of access: http://www.oucs.ox.ac.uk/rts/xaira/. – Date of access: 25.03.2012.
11. Поиск биграмм [Электронный ресурс] / А.Н. Аверин; АОТ.ru.– Режим доступа: http://www.aot.ru/demo/bigrams.html.– Дата доступа: 25.03.2012.
12. Mystem: парсер [Электронный ресурс].– Режим доступа: http://company.yandex.ru/technologies/mystem/.– Дата доступа: 25.03.2012.
13. CES: [Corpus Encoding Standard](http://www.cs.vassar.edu/CES/) [Electronic resource]. – Mode of access: http://www.cs.vassar.edu/CES/. – Date of access: 25.03.2012.
14. EAGLES: Expert Advisory Group on Language Engineering Standards [Electronic resource]. – Mode of access: http://www.ilc.cnr.it/EAGLES96/browse.html. – Date of access: 25.03.2012.
15. TEI P4: Text Encoding Initiative: Guidelines for Electronic Text Encoding and Interchange. 2001 [Electronic resource] / Sperberg-McQueen C. M., Burnard L. (eds.). – Mode of access: <http://www.hcu.ox.ac.uk/TEI/P4X/index.html>. – Date of access: 25.03.2012.
16. TEI ROMA [Electronic resource] /Arno Mittelbach; Sebastian Rahtz; Ioan Bernevig; TEI.– Mode of access: http://www.tei-c.org/Roma/.– Date of access: 25.03.2012.
17. XML Corpus Encoding Standard Document XCES 0.2 [Electronic resource]. – Mode of access: http://www.xces.org/. – Date of access: 25.03.2012.
18. Бернард, Л. Введение в TEI Lite [Электронный ресурс] / Л. Бернард, С.М. Шперберг–Мак-Куин. – Режим доступа: http://www.tei-c.org/Guidelines/Customization/Lite/teiu5\_ru.rtf. – Дата доступа: 25.03.2012.
19. ДИАЛОГ: Международная конференция по компьютерной лингвистике [Электронный ресурс] / Диалог 2000–2010.– Режим доступа: http://www.dialog-21.ru.– Дата доступа: 25.03.2012.
20. Национальный корпус русского языка и преподавание [Электронный ресурс] / Информационный ресурс по использованию корпусов в образовании.– Режим доступа: http://studiorum.ruscorpora.ru/. – Дата доступа: 25.03.2012.

**4.2. Критерии оценок результатов учебной деятельности**

|  |  |
| --- | --- |
| **Оценка** | **Критерии оценки уровня знаний и компетенций студента** |
| ***10 баллов*** | Студент дал исчерпывающий ответ на теоретический вопрос, продемонстрировал свободное владение понятийно-терминологическим аппаратом по теме вопроса билета, верно выполнил практическое задание, объяснив алгоритм его выполнения; успешно ответил на дополнительные вопросы преподавателя по программе дисциплины. |
| ***9 баллов*** | Студент дал полный ответ на теоретический вопрос, продемонстрировал владение понятийно-терминологическим аппаратом по теме вопроса билета, верно выполнил практическое задание, успешно ответил на дополнительные вопросы преподавателя по теме, смежной с вопросом / заданием билета.. |
| ***8 баллов*** | Студент дал полный ответ на теоретический вопрос, но допустил незначительные неточности, которые в дальнейшем смог устранить, отвечая на уточняющие вопросы преподавателя, с незначительными погрешностями выполнил практическое задание, но смог внести необходимые исправления по требованию преподавателя. |
| ***7 баллов*** | Студент дал полный, но с незначительными неточностями, ответ на теоретический вопрос, с незначительными погрешностями выполнил практическое задание, и при этом не смог внести необходимые исправления по требованию преподавателя. |
| ***6 баллов*** | Студент дал неполный, но удовлетворительный ответ на теоретический вопрос, с существенными погрешностями выполнил практическое задание, но смог частично внести необходимые исправления по требованию преподавателя. |
| ***5 баллов*** | Студент дал неполный, но удовлетворительный ответ на теоретический вопрос, с существенными погрешностями выполнил практическое задание, и при этом не смог внести необходимые исправления по требованию преподавателя. |
| ***4 балла*** | Студент дал неполный, но удовлетворительный ответ на теоретический вопрос, не выполнил практическое задание, но показал понимание сути задания, отвечая на наводящие вопросы преподавателя. |
| ***3 балла*** | Студент дал неудовлетворительный ответ на теоретический вопрос, не выполнил практическое задание либо выполнил с существенными погрешностями, которые не смог исправить, несмотря на наводящие вопросы преподавателя. |
| ***2 балла*** | Студент не дал ответа на теоретический вопрос, не выполнил практическое задание либо выполнил с существенными погрешностями, которые не смог исправить, несмотря на наводящие вопросы преподавателя. |
| ***1 балл*** | Студент не дал ответа на теоретический вопрос, не выполнил практическое задание либо отказался отвечать. |

Условием допуска к экзамену является успешное выполнение проекта, выполняемого студентом самостоятельно, во внеаудиторное время.

**Примерный список вопросов к экзамену**

1. Корпус как особый лингвистический ресурс.
2. Структура корпуса как лингвистического ресурса.
3. Понятие корпус-менеджера. Функции корпус-менеджера.
4. Типология лингвистических корпусов.
5. Технология создания корпуса.
6. Разметка как научная деятельность. Виды разметки.
7. Языковые и инструментальные средства разметки.
8. Разметки ручная и автоматическая, их преимущества и недостатки.
9. Корпусные технологии в компьютерной лексикографии.
10. Корпусные технологии в лингводидактике.
11. Параллельный корпус как вид лингвистического корпуса.
12. Корпус текстов обучающихся как вид лингвистического корпуса.
13. Исторический корпус как вид лингвистического корпуса.
14. Открытые корпусные проекты: OpenCorpora.
15. Открытые корпусные проекты: SPb EFL Learner Corpus.
16. Web как корпус.
17. Лингвистический корпус vs Интернет-ИПС, vs электронная библиотека, vs ресурсы типа Web as Corp.
18. Стандартизация в корпусной лингвистике.
19. TEI LIGHT как подмножество системы кодирования текстов TEI.
20. Корпусы языка специальности, доступные on-line.

**Примерная тематика проектов**

* Сравнительный анализ корпус-менеджеров (по выбору).
* Сравнительный анализ корпусов языка специальности.
* Составление мини-корпуса.
* Пополнение открытого лингвистического корпуса.
* Стандартные и дополнительные способы задания запросов к одному из лингвистических корпусов языка специальности.
* Варианты обработки средствами сторонних программ результатов выдачи к одному из корпусов языка специальности.

**СОДЕРЖАНИЕ**

**Пояснительная записка**...…………………………………….….………………………3

**Примерный тематический план** ….…………………….……………………………...5

**Содержание учебного материала**………………………………………….....................6

**Информационно-методическая часть**……………………………………..……...........8

**Критерии оценок результатов учебной деятельности...**………………………..…...11

**Примерный список вопросов к экзамену** ………....……………………………….....12

**Примерная тематика проектов**…………..…………………………………………….13

**Сведения об авторах:**

1. Рычкова Людмила Васильевна – заведующий кафедрой общего и славянского языкознания учреждения образования «Гродненский государственный университет имени Янки Купалы», кандидат филологических наук, доцент.   
Тел. +37529 8895828 (МТС).

2. Станкевич Алеся Юрьевна – старший преподаватель кафедры общего и славянского языкознания учреждения образования «Гродненский государственный университет имени Янки Купалы».