Белорусский государственный университет

УТВЕРЖДАЮ
Декан филологического ф-та профессор

_____ И.С. Ровдо
(подпись)

_____
(дата утверждения)
Регистрационный № УД-_____/р.

Спецкурс «Корпусные исследования»

Учебная программа для специальностей: 1 – 21 05 01 «Белорусская филология», 1 – 21 05 02 «Русская филология», 1 – 21 05 04 «Славянская филология», 1 – 21 05 05 «Классическая филология», 1 – 21 05 06 «Романо-германская филология», 1 – 21 05 07 «Восточная филология»

Факультет филологический_____

Кафедра прикладной лингвистики _____

Курс (курсы) 4 _____

Семестр (семестры) 7_____

Лекции __12_____ (количество часов)

Экзамен _____ (семестр)

Практические (семинарские) занятия ____16_____ (количество часов)

Зачет _____7 _____ (семестр)

Лабораторные занятия (КСР) _6__ (количество часов)

Курсовой проект (работа) _____ (семестр)

Всего аудиторных часов по дисциплине 34_____ (количество часов)

Всего часов по дисциплине _51__ (количество часов)

Форма получения высшего образования __очная_____

2009 г.

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная программа спецкурса «Корпусная лингвистика» соответствует требованиям по содержанию специализаций «Русский язык как иностранный» и «Компьютерная лингвистика». В программе представлен и обобщен опыт преподавания данной дисциплины на кафедре прикладной лингвистики филологического факультета Белгосуниверситета. Данная учебная программа основывается на разделе «Корпусная лингвистика» учебной программы «Компьютерная лингвистика».

Магистрантам предлагается рассмотрение проблем создания и использования языковых корпусов. При этом учитываются современные методики создания корпусов, опыт создания корпусов за рубежом и в нашей стране, прагматические аспекты корпусной лингвистики.

Целями данного курса являются:

– Ознакомление студентов с новой парадигмой в лингвистических исследованиях;
– Ознакомление студентов с историей корпусных исследований;
– Изучение языковых и программных средств корпусной лингвистики;
– Формирование навыков работы с программными средствами и информационными ресурсами корпусной лингвистики.

Реализация заявленных целей опирается на решение следующих задач:

- расширение языковой компетенции иностранных учащихся на основе систематизации знаний о языке и его коммуникативных возможностях;
- определение специфики русского языка с точки зрения проблемы взаимосвязи языка и интеллекта.

Данная программа является частью программы по современному русскому языку (научный стиль речи).

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

1. Понятие о корпусной лингвистике.

Корпусная лингвистика — раздел языкознания, занимающийся разработкой, созданием и использованием текстовых (лингвистических) корпусов. Термин введён в употребление в 60-х годах XX века в связи с развитием практики создания корпусов, которому начиная с 80-х способствовало развитие вычислительной техники.

Лингвистическим корпусом называют собрание текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой. Иногда корпусом («корпус первого порядка») называют просто любое собрание текстов, объединённых каким-то общим признаком (языком, жанром, автором, периодом создания текстов).

Целесообразность создания текстовых корпусов объясняется:

- представлением лингвистических данных в реальном контексте;
- достаточно большой представительностью данных (при большом объёме корпуса);
- возможностью многократного использования единожды созданного корпуса для решения различных лингвистических задач.

2. История корпусной лингвистики.

Первым большим компьютерным корпусом считается Брауновский корпус (БК, англ. Brown Corpus, BC), который был создан в 1960-е годов в Университете Брауна и содержал 500 фрагментов текстов по 2 тысячи слов в каждом, которые были опубликованы на английском языке в США в 1961 году. В результате он задал стандарт в 1 млн словоупотреблений для создания представительных корпусов на других языках. По модели близкой к БК в 1970-е годы был создан частотный словарь русского языка Засориной, построенный на основе корпуса текстов объемом также в 1 миллион слов и включавший примерно в равной пропорции общественно-политические тексты, художественную литературу, научные и научно-популярные тексты из разных областей и драматургию. По аналогичной модели был построен и русский корпус, созданный в 1980-е годы в Университете Уппсалы, Швеция.

Размер в один миллион слов достаточен для лексикографического описания только самых частотных слов, поскольку слова и грамматические конструкции средней частоты встречаются по несколько раз на миллион слов (со статистической точки зрения язык является большим набором редких событий). Так каждое из таких обыденных слов англ. polite (вежливый) или англ. sunshine (солнечный свет) встречается в БК всего 7 раз, выражение англ. polite letter лишь один раз, а такие устойчивые выражения как англ. polite conversation, smile, request ни разу.

По этим причинам, а также в связи с ростом компьютерных мощностей, способных работать с большими объемами текстов, в 1980-е годы в мире было предпринято несколько попыток создать корпуса большего размера. В Великобритании такими проектами были Банк Английского (Bank of English) и Британский Национальный Корпус (British

National Corpus, BNC). В СССР таким проектом был Машинный Фонд русского языка, создававшийся по инициативе А. П. Ершова.

3. Современное состояние корпусной лингвистики.

Наличие большого количества текстов в электронной форме существенно облегчило задачу создания больших представительных корпусов размером в десятки и сотни миллионов слов, но не ликвидировало проблем: сбор тысяч текстов, снятие проблем с авторскими правами, приведение всех текстов в единую форму, балансировка корпуса по темам и жанрам отнимают много времени. Представительные корпуса существуют (или разрабатываются) для немецкого, польского, чешского, словенского, финского, новогреческого, армянского, китайского, японского и других языков.

Национальный корпус русского языка, создаваемый при РАН, содержит на сегодняшний день более 140 млн словоупотреблений.

Наряду с представительными корпусами, которые охватывают большой набор жанров и функциональных стилей, в лингвистических исследованиях часто используются и оппортунистические коллекции текстов, например, газеты (часто Wall Street Journal и New York Times), новостные ленты (Рейтер), коллекции художественной литературы (Библиотека Мошкова или Проект Гутенберг).

4. Проблемы корпусной лингвистики.

Корпус состоит из конечного числа текстов, но он призван адекватно отражать лексикограмматические феномены, типичные для всего объема текстов в соответствующем языке (или подъязыке). Для представительности важен как размер, так и структура корпуса. Представительный размер зависит от задачи, поскольку он определяется тем, как много примеров может быть найдено для исследуемых феноменов. В связи с тем, что со статистической точки зрения язык содержит большое число относительно редких слов (Закон Ципфа), для исследования первых пяти тысяч наиболее частотных слов (например, убыток, извиняться) требуется корпус размером около 10-20 миллионов словоупотреблений, в то время как для описания первых двадцати тысяч слов (незатейливый, сердцебиение, роиться) уже требуется корпус свыше ста миллионов словоупотреблений.

К первичной разметке текстов относятся этапы, обязательные для каждого корпуса:

- токенизация (разбиение на орфографические слова)
- лемматизация (приведение словоформ к словарной форме)
- морфологический анализ.

В больших корпусах возникает проблема, которая ранее была неактуальной: поиск по запросу может выдавать сотни и даже тысячи результатов (контекстов употребления), которые просто физически невозможно просмотреть в ограниченное время. Для решения этой проблемы разрабатываются системы, позволяющие группировать результаты поиска и автоматически разбивать их на подмножества (кластеризация результатов поиска), либо выдающие наиболее устойчивые словосочетания (коллокации) со статистической оценкой их значимости.

5. Веб как корпус.

В качестве корпуса может использоваться множество текстов, доступных в интернете (то есть миллиарды словоупотреблений для основных мировых языков). Для лингвистов самым распространенным способом работы с Интернетом остаётся составление запросов к поисковой машине и интерпретация результатов либо по числу найденных страниц, либо по первым возвращенным ссылкам. В английском языке такая методология получила название англ. Googleology, для русского более подходящим названием может стать Яндексология. Необходимо отметить, что такой подход годится для решения ограниченного класса задач, так как средства разметки текстов, используемые в вебе, не описывают ряд лингвистических особенностей текста (указание ударений, грамматических классов, границ словосочетаний и т. д.). Кроме того дело осложняется малой распространённостью семантической вёрстки.

На практике ограниченность такого подхода приводит к тому, что проверить, например, сочетаемость двух слов проще всего через запрос вида «слово1 слово2». По полученным результатам можно судить, насколько распространено такое сочетание и в каких текстах оно чаще встречается.

Второй способ заключается в автоматическом извлечении большого количества страниц из Интернета и их дальнейшем использовании в качестве обычного корпуса, что дает возможность провести его разметку и использовать лингвистические параметры в запросах. Этот способ позволяет быстро создать представительный корпус для любого языка в достаточной степени представленного в Интернете, но его жанровое и тематическое разнообразие будет отражать интересы пользователей Интернета.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА

Номер раздела, тем	Название раздела, темы, занятия; перечень изучаемых вопросов	Количество аудиторных часов	Материальное обеспечение занятия (наглядные,	Литература

ы, занятия		лекции				методические пособия и др.)	
1	2	3	4	5	6	7	8
1.	Корпус текстов, его особенности. Виды корпусов текстов.	2				Раздаточ-ный материал, слайды	[2], [15]
2.	Наиважнейшие компьтерные корпусы текстов и возможности их использования;		4			Раздаточ-ный материал, слайды	[2], [15], [21]
3.	Корпусная лингвистика, её значение, краткая история. Корпусные исследования в Беларуси.	2				Раздаточ-ный материал	[9], [16]
4.	Основные понятия и принципы корпусной лингвистики. Корпус текстов как условие объективного лингвистического исследования.	2				Раздаточ-ный материал	[2]
5.	Лингвистически аннотированные корпуса русского языка (Обзор общедоступных ресурсов).		4			Раздаточ-ный материал, слайды	[1], [6], [10], [11]
6.	Сопоставление корпусной и традиционной лингвистики.	2				Раздаточ-ный материал, слайды	[2], [15]
7.	Корпус как особый тип информационно-поисковой системы.	4				Раздаточ-ный материал, слайды	[8]
8.	Экстралингвистическая разметка. Метаданные.			2		Раздаточ-ный материал, слайды	[2], [8]
9.	Лингвистическая разметка, её виды.			2		Раздаточ-ный материал, слайды	[8], [18]
10.	Корпусные менеджеры.		4			Раздаточ-ный материал,	[8], [9]

						слайды		
11.	Использование корпусов в прикладной лингвистике и других областях.		4			Раздаточ-ный материал, слайды	[8]	
12.	Технология создания корпусов.			2		Раздаточ-ный материал	[2]	
	ВСЕГО	**12**	**16**	**6**				

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

1. Андрющенко В.М. Концепция и архитектура машинного фонда русского языка / Отв. ред. А.П. Ершов. М., 1989.
2. Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику. М., 2001. С. 112–137.
3. Богуславский И.М. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды международного семинара по компьютерной лингвистике и ее приложениям «Диалог 2000». Протвино, 2000.
4. Венцов А.В., Касевич В.Б., Ягунова Е.В. Корпус русского языка и восприятие речи // НТИ. Сер. 2. 2003. № 6. С. 25–32.
5. Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // НТИ. Сер. 2. 2003. № 6. С. 2–8.
6. Добровольский Д. О. Корпус параллельных текстов как инструмент анализа литературного перевода //http://www.dialog-21.ru/Archive/2003/Dobrovolskij.htm.
7. Доклады научной конференции «Корпусная лингвистика лингвистические базы данных» / Под ред. А.С. Герда. СПб., 2002.
8. Захаров В.П. Информационные системы (документальный поиск). Санкт-Петербург, 2002.
9. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005.
10. Захаров В.П. Чешский национальный корпус текстов: организация и способы использования // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под ред. А.С. Герда. СПб., 2002. С. 72–79.
11. Копотев М.В., Мустайоки А. Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет // НТИ. Сер. 2. 2003. № 6. С. 33–36. 5. Копотев М.В. Корпусная лингвистика в Финляндии (обзор ресурсов) // НТИ. Сер. 2. 2003. № 6. С. 37–41.
12. Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005.
13. Научно-техническая информация. Сер. 2. 2005. № 3.
14. Научно-техническая информация. Сер. 2. 2003. № 6.
15. Рыков В.В. Прагматически ориентированный корпус текстов // Тверской лингвистический меридиан. Вып. 3. Тверь, 1999. С. 89–96.
16. Труды международного семинара по компьютерной лингвистике и ее приложениям «Диалог 2000», «Диалог 2001», «Диалог 2002», «Диалог 2003», «Диалог 2004», «Диалог 2005».
17. Труды международной научной конференции «Корпусная лингвистика 2004» / Под ред. А.С. Герда. СПб., 2004.
18. Чардин И.С. Лингвистические корпуса с синтаксической разметкой и их применение // НТИ. Сер. 2. 2003. № 6. С. 18–24.
19. Шаров С.А. Параметры описания текстов корпуса.
20. Шаров С.А. Формат выходного представления корпуса текстов.
21. Шаров С.А. Представительный корпус русского языка в контексте

мирового опыта // НТИ. Сер. 2. 2003. № 6. С. 9–17.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ С ДРУГИМИ ДИСЦИПЛИНАМИ СПЕЦИАЛЬНОСТИ

Название дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы по изучаемой учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
1. Современный русский язык	Кафедра прикладной лингвистики		
2. Методика преподавания РКИ	Кафедра прикладной лингвистики		
3. Общее языкознание	Кафедра общего и теоретического языкознания		

Данная учебная программа является частью следующей учебной программы: В. А. Карпов, А. В. Лаврененко, А. И. Головня. Компьютерная лингвистика. Учебная программа по специализации «Компьютерная лингвистика»: для студентов специальностей 1-21 05 01 «Белорусская филология», 1-21 05 02 «Русская филология», 1-21 05 04 «Славянская

филология», 1-21 05 05 «Классическая филология», 1-21 05 06 «Романо-германская филология», 1-21 05 07 «Восточная филология» (Рекомендовано Учёным советом филологического факультета 25 октября 2008, протокол № 2). – Минск, 2008.

Рассмотрена и рекомендована к утверждению на заседании кафедры прикладной лингвистики_____ _____
(дата, номер протокола)

Заведующий кафедрой
_____ Л.Ф. Гербик
(подпись)

Одобрена и рекомендована к утверждению Научно-методическим советом филологического факультета Белгосуниверситета_____
(дата, номер протокола)

Председатель
_____ _____
(подпись) (И.О.Фамилия)

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная программа спецкурса «Корпусные исследования» соответствует требованиям по содержанию специализации «Русский язык как иностранный». В программе представлен и обобщен опыт преподавания данной дисциплины на кафедре прикладной лингвистики филологического факультета Белгосуниверситета. Данная учебная программа основывается на

разделе «Корпусная лингвистика» учебной программы «Компьютерная лингвистика».

Студентам предлагается рассмотрение проблем создания и использования языковых корпусов. При этом учитываются современные методики создания корпусов, опыт создания корпусов за рубежом и в нашей стране, прагматические аспекты корпусной лингвистики.

Целями данного курса являются:

– Ознакомление студентов с новой парадигмой в лингвистических исследованиях;
– Ознакомление студентов с историей корпусных исследований;
– Изучение языковых и программных средств корпусной лингвистики;
– Формирование навыков работы с программными средствами и информационными ресурсами корпусной лингвистики.

Реализация заявленных целей опирается на решение следующих задач:

- расширение языковой компетенции иностранных учащихся на основе систематизации знаний о языке и его коммуникативных возможностях;
- определение специфики русского языка с точки зрения проблемы взаимосвязи языка и интеллекта.

Белорусский государственный университет

УТВЕРЖДАЮ
Декан филологического ф-та профессор

_____ И.С. Ровдо
(подпись)

_____
(дата утверждения)

Регистрационный № УД-_____/р.

Спецсеминар
«Корпусная лингвистика»
Учебная программа для специальностей:
1 – 21 05 01 «Белорусская филология», 1 – 21 05 02 «Русская филология», 1 – 21 05 04 «Славянская филология», 1 – 21 05 05 «Классическая филология», 1 – 21 05 06 «Романо-германская филология», 1 – 21 05 07 «Восточная филология»

Факультет филологический_____

Кафедра прикладной лингвистики _____

Курс (курсы) 3 – 5 _____

Семестр (семестры) 5 – 9 _____

Лекции __24_____ (количество часов) Экзамен _____ (семестр)

Практические (семинарские) занятия ____32_____ (количество часов) Зачет _____5 – 9_____ (семестр)

Лабораторные занятия (КСР) _12__ (количество часов) Курсовой проект (работа) _____ (семестр)

Всего аудиторных часов по дисциплине 68_____ (количество часов)

Всего часов по дисциплине _102____ (количество часов) Форма получения высшего образования __очная_____

2009 г.