Министерство образования и науки РФ
Федеральное государственное автономное образовательное учреждение высшего образования
«Новосибирский национальный исследовательский государственный университет» (Новосибирский государственный университет, НГУ)
Гуманитарный факультет

Программа рассмотрена
на заседании кафедры
фундаментальной и прикладной
лингвистики
29.08.2014
_____
Зав. кафедрой, проф. М.К. Тимофеева

Утверждаю

_____
декан гуманитарного
факультета, профессор
1.09.2014
Л.Г. Панин

Основная образовательная программа
высшего образования

Направление подготовки
035800 – Фундаментальная и прикладная лингвистика

Квалификация (степень) выпускника –
бакалавр

ПРОГРАММА УЧЕБНОГО КУРСА «КОРПУСНАЯ ЛИНГВИСТИКА»

(76 часов, 2 з.е.)

1. **Наименование дисциплины**

ПРОГРАММА УЧЕБНОГО КУРСА
«КОРПУСНАЯ ЛИНГВИСТИКА»

Программа дисциплины «Корпусная лингвистика» составлена в соответствии с требованиями к обязательному минимуму содержания и уровню подготовки дипломированного бакалавра по направлению 035800 «Фундаментальная и прикладная лингвистика» в целях обеспечения реализации учебного процесса в НГУ.

Автор: Алешина Ольга Николаевна, доктор филол. наук, профессор

2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы.

Цели освоения специального курса

Дисциплина (курс) «Корпусная лингвистика» имеет своими целями:

1) познакомить обучающихся с основными понятиями корпусной лингвистики;

2) научить их использовать современные российские и зарубежные корпусные собрания в достижении собственных исследовательских задач.

В результате освоения дисциплины обучающийся должен:

- Знать базовую терминологию корпусной лингвистики и типологию российских и зарубежных лингвистических корпусов; иметь сведения о технологических этапах и целях их создания.
- Уметь выбрать и правильно применять корпусные ресурсы в решении собственных исследовательских задач.
- Владеть навыками составления и статистической обработки индивидуальной исследовательской картотеки, созданной на основе существующих лингвистических корпусов.

Перечисленные результаты образования являются основой для формирования следующих общекультурных и общепрофессиональных компетенций:

***а) общекультурными* (*ОК*)**

- умением логически верно, аргументировано и ясно строить устную и письменную речь — ОК-2
- использованием основных положений и методов социальных, гуманитарных и экономических наук при решении социальных и профессиональных задач, способностью анализировать социально- значимые проблемы и процессы — ОК-9

***б) профессиональными (ПК)*:**

общепрофессиональными:

- знанием основных понятий и категорий современной лингвистики — ПК-1

в области научно-исследовательской деятельности:

– владением основными способами описания и формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации, содержащейся в тексте на естественном языке — ПК-11

3. Место дисциплины в структуре образовательной программы

Изучение курса «Корпусная лингвистика» возможно после освоения программы базового курса «Современный русский язык».

4. Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 76 часов, 2 з.е., 32 часа лекции, 40 часа самостоятельная работа.

№ п/п	Раздел Дисциплины	Семестр	Неделя семестра	Виды учебной работы, включая самостоятельную работу студентов и трудоемкость (в часах)				Формы текущего контроля успеваемости *(по неделям семестра)* Форма промежуточной аттестации *(по семестрам)*
				Лекции	Практикум	Самостоят. Работа	Всего	
1	*Тема 1*. Базовые понятия и термины корпусной лингвистики.			2	2	3	7	
2	*Тема 2*. Компьютерный лингвистический корпус как база текстовых данных.			2	2	3	7	
3	*Тема 3*. Технологический процесс создания компьютерного лингвистического корпуса.			1	1	3	5	
4	*Тема 4*. «Проблемные зоны» современной русской корпусной лингвистики, о которых должен знать пользователь.			1	1	3	5	
5	*Тема 5*. Корпусная типология. Национальные лингвистические корпусы.			2	1	4	7	
6	*Тема 6*. Компьютерная лексикография. Компьютерные тезаурусы RusNet и др.			2	2	4	8	
7	*Тема 7*. Авторская компьютерная лексикография.			1	1	3	5	

8	*Тема 8*. Компьютерные корпусы устной речи.			2	2	3	7	
9	*Тема 9*. Корпусная лингвистика и проблемы перевода.			1	1	3	5	
10	*Тема 10*. Многоязычные ресурсы и параллельные корпусы.			1	2	2	5	
11	*Тема 11*. Жанрово ограниченные корпусы и возможности привлечения их ресурсов в прикладных целях.			1	1	3	5	
12	*Тема 12*. Исторические корпусы.			1	1	3	5	
13	*Тема 13*. «Библометрические» возможности применения корпусных собраний.			1	1	3	5	
	Всего			18	18	40	76	Итоговый зачет

Тема 1. Базовые понятия и термины корпусной лингвистики.

Объект и предмет корпусной лингвистики. Корпусная лингвистика как раздел компьютерной лингвистики, занимающийся на основе привлечения компьютерных технологий разработкой принципов: 1) создания лингвистических корпусов и 2) применения созданных корпусов в филологических («библометрических» и др.) исследованиях.

Корпус и компьютерный корпус (J. Sinclair). Содержание понятия «компьютерный корпус» как базового понятия корпусной лингвистики.

Компьютерный корпус как объект лингвистического аннотирования. Корпусная аннотация, или корпусная разметка (англ. annotation, tagging), включающая: а) метаразметку как приписывание текстам экспралингвистических меток (англ. tags) - сведений об их авторах, в том числе биографических, историко-библиографической информации, данных о жанрово-тематической отнесенности текстов; б) структурную разметку как маркирование текстов по их структурным частям (главам, абзацам, фразам, текстоформам и словоформам); в) собственно лингвистическую разметку на основе лемматизации (просодическую (и дискурсивную); семантическую; морфологическую, или частеречную (англ. part-of-speech tagging, или POS-tagging); анафорическую (маркирование средств выражения дейксиса и их кореферентов); синтаксическую, или парсинг (англ. parsing). Проблемы терминоупотребления в отечественной корпусной лингвистике.

Корпусный менеджер (корпус-менеджер, англ. corpus manager), или программа-конкордансер (англ. concordancer), как система с программными средствами, применяемая для поиска данных в электронных версиях текстов, их статистической обработки и предоставления результатов пользователям. Конкордансеры WordSmith Tools, AntConc, Dialing Concordance и др.

Глобальные информационно-поисковые системы вербального типа как корпусные менеджеры (Google, AltaVista, Яндекс, Rambler и др.). Примеры их использования в лингвистических исследованиях.

Тема 2. Компьютерный лингвистический корпус как база текстовых данных.

Компьютерный корпус как объемная, представленная в электронном виде, структурно унифицированная и последовательно размеченная специалистами база текстовых данных. Различия между электронным архивом, электронной библиотекой и лингвистическим корпусом. Закон Ципфа (G. K. Zipf) и вопрос о репрезентативности объема корпусного материала для решения частных лингвистических задач.

Понятия «мегакорпус» (на примерах BNC – British National Corpus, COCA – Corpus of Contemporary American English и др.) и «подкорпус», или «субкорпус» (на примере подкорпусов НКРЯ – Национального корпуса русского языка).

Сопоставление объектных и предметных областей «традиционной» и корпусной лингвистики (по В.В. Рыкову и др.). Зависимость исследовательских результатов «итологии» и «библометрии» от уровня развития IT-технологий и технологических мощностей компьютерного оборудования. Краткий анализ отечественных и зарубежных периодических специализированных изданий по корпусной лингвистике. Центры преподавания корпусной лингвистики в России и за рубежом.

Тема 3. Технологический процесс создания компьютерного лингвистического корпуса.

Общая характеристика лингвистических информационно-программных «изделий». Технологические этапы создания лингвистического корпуса: 1) формулировка цели создания корпуса; 2) определение круга текстовых источников; 3) создание электронных вариантов выбранных текстов, то есть их оцифровка, или «ввод в компьютер» с помощью набора (для устных и письменных текстов) и/или сканирования (для письменных текстов) с обязательной последующей сверкой электронных копий с оригиналами; 4) аннотирование: а) паспортизация текстов, или их метаразметка; б) их конвертирование в текстовый режим с последующей структурной разметкой; в) лингвистическое аннотирование; 5) конвертирование размеченных текстов в корпусный менеджер; 6) обеспечение доступа пользователям к корпусным ресурсам.

Использование при создании корпусов программ автоматического морфологического и синтаксического аннотирования – тэггеров (англ. taggers) и парсеров (англ. parsers). Понятия «лемма», «лемматизация», «хэширование» (англ. hash). Трудности лемматизации, обусловленные: 1) трудоемкостью процесса разметки вручную; 2) отсутствием полного, глубокого, всестороннего и непротиворечивого теоретического лингвистического описания естественных языков, на которых созданы тексты; 3) непоследовательностью структурной и семантической организации всех подсистем естественных языков.

Тема 4. «Проблемные зоны» современной русской корпусной лингвистики, о которых должен знать пользователь.

Предоставляемые пользователям корпусных собраний возможности: 1) в сборе исследовательской картотеки – а) поиск репрезентантов заранее заданных семантических и / или грамматических категорий; б) поиск словоформ и текстоформ по леммам; в) поиск устойчивых сочетаний; г) поиск текстов с заданными металингвистическими параметрами; 2) в анализе собранной информации – а) составление конкордансов с целью уточнения парадигматических и синтагматических связей языковых единиц; б) получение статистических данных о языковых и текстовых единицах и связанных с ними категориях; 3) эмпирически достоверное подтверждение или опровержение исследовательской гипотезы.

Проблема унификации и последующей стандартизации аннотирования. Трудности приведения мета- и собственно лингвистической разметки текстов в соответствие со стандартами Text Encoding Initiative (TEI) и Expert Advisory Group on Language Engineering Standards (EAGLES). Понятие формального языка разметки (языки SGML и XML). «Постулаты аннотирования» Дж. Лича (G. Leech) и их «нарушение».Сложности, связанные с конкордансным характером подачи текстовой информации в корпусных собраниях. Проблема кластеризации результатов поиска.

Тема 5. Корпусная типология. Национальные лингвистические корпусы.

Краткая история создания зарубежных и отечественных компьютерных корпусных собраний.

Корпусы «первого поколения»: Brown Corpus, WSC – Wellington Corpus of Spoken New Zealand English, Машинный фонд русского языка и др. Корпусы «второго поколения»: BNC, COCA, Bank of English, ХАНКО – Хельсинкский аннотированный корпус русских текстов, НКРЯ и др.

Корпусная типология (на основе типологии В. П. Захарова и др.): 1) по типу речевых данных (письменные (текстовые), устные (дискурсивные) и смешанные); 2) по языку (русские, английские, французские, китайские и т.д.); 3) по наличию / отсутствию переводческих «параллелей» (одноязычные, двуязычные, многоязычные); 4) по социолектной спецификации текстов (литературные (англ. standart), нелитературные (диалектные, частные социолектные) и смешанные); 5) по жанровой спецификации текстов (прозаические (фольклорные, художественные и / или публицистические); поэтические (фольклорные, художественные и / или публицистические); смешанные); 6) по доступности для пользователей (существующие в открытом доступе, коммерческие, закрытые); 7) по целевому назначению (исследовательские и иллюстративные); 8) по «динамичности» пополнения состава корпуса («динамические» (в том числе «мониторные») и «статические»); 9) по наличию / отсутствию мета- и собственно текстовой разметки (аннотированные и неаннотированные); 10) по характеру разметки (метааннотированные, структурно аннотированные и лингвистически (семантически, просодически, морфологически, синтаксически и т.д.) аннотированные); 11) по полноте объема текстов собрания (полнотекстовые и «фрагментнотекстовые», или неполнотекстовые); 12) по временной атрибуции текстов (современные и исторические); 13) по количеству произведений и

их авторской атрибуции (неанонимные коллективные / идиолектные и анонимные коллективные / идиолектные); и т.д. Примеры корпусных проектов каждого типа.

Общая характеристика лингвистических корпусов наиболее распространенных языков мира: 1) китайского (Корпус китайского языка, создаваемый сотрудниками Пекинского университета, и др.; 2) английского (Bank of English и его Британский и Американский подкорпусы: источники текстовых собраний, специфика частеречного аннотирования (на фоне традиционных грамматических описаний) и возможности лингвистической интерпретации корпусных ресурсов); 3) арабского (CCA – Corpus of Contemporary Arabic); 4) испанского (Corpus del Español); 5) русского (НКРЯ). Представление результатов аннотирования в национальных корпусах (на примерах BNC и НКРЯ).

Тема 6. Компьютерная лексикография. Компьютерные тезаурусы RusNet и др.

Краткая характеристика основ идеографического описания языков (на примерах «Ономастикона» Поллукса, «Амаракоша», симфоний Библии, трудов Д. Дальгарно, Дж. Уилкинса и современных исследователей). Понятие «семантической сети» как одного из способов отражения семантики предметной области и как промежуточной формы представления текстовых данных. Классификации семантических сетей и их критика. Типы отношений в семантических сетях. Понятия «гипертекст» и «семантическая паутина». Тезаурус как особая разновидность словарей (на примере Roget's Thesaurus). «Русский семантический словарь» как опыт современного системного идеографического представления семантики.

Стандартные методики построения компьютерных словарей. Источники материала для wordnet-словарей. Перечисление значений слов в соответствии с их частотным распределением в корпусе текстов как один из базовых принципов построения wordnet-тезауруса. Представление «синсетов» и сочетаемостных данных в компьютерных словарях. Проекты WordNet, EuroWordNet, BalkaNet. Цель создания компьютерного тезауруса RussNet. Проект ThIE – Индоевропейского компьютерного тезауруса. Гипертекстовый тезаурус ЛСВ русского языка.

Тема 7. Авторская компьютерная лексикография.

Идеология создания писательских словарей. Предоставление исчерпывающей и объективной информации о языке писателя как основная цель авторской лексикографии. Проблемы традиционного авторского языкового «портретирования»: сложность и трудоемкость сбора картотеки, фиксация отдельных примеров в качестве иллюстраторов значений, исследовательский субъективизм в разделении значений многозначных слов, ограниченный объем подачи словарной информации в печатном издании. Создание идиолектных корпусов как перспективное направление в авторской лексикографии.

Конкордансный проект «Весь Достоевский», разработанный сотрудниками Петрозаводского государственного университета и позволяющий получить контекст употребления и частотные ха-

рактеристики слов и словоформ, употребленных писателем. «Электронная *энциклопедия* языка Пушкина», создаваемая сотрудниками кафедры русского языка и лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ.

«Словарь-корпус языка А.С. Грибоедова» как первое полное описание языка русского писателя. Алфавитно-частотный конкорданс как основная часть словаря; вспомогательные словари и указатели (алфавитный указатель к тексту «Горя от ума»; частотный словарь; частотные словари прозаической и поэтической речи; обратный алфавитный словарь; грамматический словарь). Структура словарной статьи: заглавное («заголовочное») слово, грамматические пометы (часть речи, вид, род, одушевленность / неодушевленность и т.д.), количественная помета, указывающая суммарную частоту по всем текстам, краткое толкование (в случаях расхождения с современным).

Общая пользовательская характеристика ГАФИС «Брюсов» – гипертекстовой авторской филологической информационной системы по творчеству В. Я. Брюсова, разрабатываемой сотрудниками кафедры теоретической лингвистики и кафедры математики, логики и интеллектуальных систем РГГУ.

Тема 8. Компьютерные корпусы устной речи.

Традиционные способы фиксации устной речи и применение их результатов в корпусной лингвистике. Создание компьютерных устных текстов как направление, открывающее новые перспективы для лингвистической теории и практики. Общая характеристика устных корпусов на примерах проектов TalkBank, BASE – British Academic Spoken English, CANCODE – Cambridge and Nottingham Corpus of Discourse in English, CSLU Speech Corpora – Center for Spoken Language Understanding, ELISA – English Language Interview Corpus as a Second-Language Application, EUSTAGE – Edinburgh University Speech Timing Archive and Corpus of English, IViE - Intonational Variation in English, MICASE – Michigan Corpus of Academic Spoken English и др., а также подкорпуса BNC.

Использование материалов корпусных собраний в диалектологии (на примерах Northern Ireland Transcribed Corpus, L-CIE - Limerick Corpus of Irish-English, Nationwide Speech Project Corpus, SBCSAE - Santa Barbara Corpus of Spoken American English, NECTE - Newcastle Electronic Corpus of Tyneside English, Spoken Corpus of the Survey of English Dialects, FRED – Freiburg English Dialect Corpus, WSC - Wellington Corpus of Spoken New Zealand English), социолингвистике (DCPSE - Diachronic Corpus of Present-Day Spoken English, COLT - Bergen Corpus of London Teenage Language, ANDOSL - Australian National Database of Spoken Language, Longman Spoken American Corpus, Northern Ireland Transcribed Corpus и др.), паралингвистике (проекты Emotional Speech Corpus, EEKK- Estonian Emotional Speech Corpus, Finnish Emotional Speech Database), языковой контактологии (LeaP - Learning the Prosody of a foreign language).

Разработка российских проектов «Мультимедийный корпус русских разговорных текстов», «Звуковой корпус русского языка повседневного общения “Один речевой день”» (ORD Speech Corpus of Russian Everyday Communication). Устный подкорпус в составе НКРЯ.

Тема 9. Корпусная лингвистика и проблемы перевода.

Лингвистическая относительность и эмпирически сложившиеся способы ее преодоления.

Два вида компьютерных переводческих технологий: 1) машинный, или автоматический (англ. MT, machine translation); 2) CAT (англ. computer-assisted/aided translation), связанный с программами Trados, OmegaT, DejaVu, WordFast и т.п., реализующими концепцию «памяти переводов» (англ «translation memory»). Обязательность обучения основам CAT в европейских и американских университетских курсах теории и практики перевода. Общая характеристика важных для лингвиста-переводчика Интернет-ресурсов. Перспективность приблизительного автоматического перевода (на примерах http://www.translate.ru/ и www.google.com/language_tools). Понятие «коллокат». Конкордансеры как программы предпереводческого анализа текстов, дающие возможность увидеть все контексты употребления определённой словоформы и коллокаты. Терминологические базы данных как программы, позволяющие переводчику работать с электронными версиями различных словарей и составлять собственные. Программа ПарТекс, ориентированная на обнаружение асимметрии в параллельных текстах.

Тема 10. Многоязычные ресурсы и параллельные корпусы.

Параллельные корпусы как средство решения вопросов теории и практики перевода. Концепции создания параллельных корпусов (на примерах: 1) Англо-русского и франко-русского параллельного тегированного корпуса, разрабатываемого в Минском государственном лингвистическом университете; 2) Англо-русского параллельного выровненного корпуса текстов, разрабатываемого членами Санкт-Петербургского отделения Союза переводчиков России). Многоязычный историко-параллельный корпус СПИ - «Слово о полку Игореве» как один из крупнейших специализированных переводческих корпусных проектов. Формирование в составе НКРЯ подкорпуса параллельных текстов (КоПарТ). Технологические и собственно лингвистические проблемы в создании устных параллельных корпусов (на примере Корпуса устных судебных переводов, создаваемого сотрудниками Тамперского университета (Финляндия)).

Тема 11. Жанрово ограниченные корпусы и возможности привлечения их ресурсов в прикладных целях.

КОРГАЗ – компьютерный корпус текстов русских газет конца XX века: общая характеристика корпусной продукции на базе текстовых файлов русскоязычных российских газет, накопленных в информационной системе фирмы Интегрум-Техно и обработанных сотрудниками лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ; системный продукт Диктум-1; объем корпуса. Получение объективной картины состояния русского газетного языка конца XX века как одна из задач составителей и разработчиков корпусного собрания и соотношение источников текстовых материалов по их объёмам в КОРГАЗ. Поня-

тие «ядерный газетный корпус». Профили распределения объемов текстов разного жанрового и источникового состава и особенностей употребления в них словоформ. Возможности создания на основе корпусных данных алфавитно-частотных и частотно-распределительных словарей словоформ, лемм, корней и морфемных моделей, реализованных в русском газетном языке конца XX века. Выявление основных газетных жанровообразующих факторов на основе статистического анализа текстов КОРГАЗ.

Газетный и художественный подкорпусы Лидских корпусов русских текстов 2000-2001 гг. (Университет Лидса, Великобритания). Возможности, предоставляемые корпусом в выявлении жанровой этноспецифики русского и английской газетного (language of newspapers) и художественного (language of fiction) языка.

Общая характеристика подкорпуса газетных интервью (с 1996 г. по н.вр.) в Упсальском корпусе современных русских текстов, разрабатываемом на отделении славистики Университета Упсалы (Швеция). «Частотный словарь современного русского языка» (проф. L. Lönngren, 1993 г.) и источники его создания.

Возможности применения корпусных ресурсов в политической лингвистике и пиар-технологиях.

Тема 12. Исторические корпусы.

Использование электронных собраний в исторической и авторской лексикографии. Принципы представления буквенно-графических составов рукописных текстов в корпусных базах данных.

Общая характеристика концепций: 1) ARCHER - A Representative Corpus of Historical English Registers (1650-2005 гг.); 2) YCOE - The York-Toronto-Helsinki Parsed Corpus of Old English Prose; 3) Корпуса старолатышских конфессиональных и юридических текстов (XVI - начало XVIII вв.), разрабатываемого сотрудниками Института математики и информатики Латвийского университета, как примера не полностью грамматически аннотированного корпуса; 4) СКАТ – Санкт-Петербургского корпуса агиографических текстов, созданного на кафедре математической лингвистики факультета филологии и искусств СПбГУ; 5) Корпуса русского языка XVIII века, создаваемого в НИЦ «Культурное наследие и информационные технологии» Казанского государственного университета.

Представление древнерусских редакций в СПИ – Историко-параллельном корпусе «Слова о полку Игореве».

Тема 13. «Библометрические» возможности применения корпусных собраний.

Привлечение лингвистических корпусов для проведения разноаспектных исследований в области: 1) фонетики, 2) лексикологии и лексикографии; 3) словообразования; 4) морфологии; 5) морфонологии; 6) фразеологии; 7) синтаксиса языков.

«Снятие» андроцентризма традиционной лексикографии в корпусных исследованиях и другие возможности применения корпусных собраний в гендерной лингвистике.

Понятие «стилометрия» («стилеметрия»). Использование корпусных данных в литературоведении и стилистике. Роль национальных и идиолектных корпусов в объективном решении филологических задач, в частности в: 1) исследованиях особенностей поэтики и стиля конкретного автора или ряда авторов определенного литературного направления, 2) стилометрическом анализе текстов поэтических и прозаических жанров, 3) исследовании употреблений стилевых приемов конкретными писателями; и т.д.

Применение корпусных материалов в ортологии и в преподавании иностранных и неродных языков (Lancaster Corpus of Children's Project Writing, ICLE - International Corpus of Learner English, CHILDES – Child Language Data Exchange System и др.). Корпусные ресурсы для журналистов и рекламоисполнителей.

5. Образовательные технологии

Для аудиторных лекционных, практических и самостоятельных занятий необходимы компьютерное обеспечение и доступ к Интернету.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины и учебно-методическое обеспечение самостоятельной работы студентов.

В ходе изучения программных тем обучаемые выполняют практические и самостоятельные работы. Зачет сдается по итогам всего курса.

Список практических работ:

1. Лингвистическое аннотирование мегакорпусов (на примерах систем разметок НКРЯ и BNC).
2. Сопоставительный анализ пользовательских возможностей русских аннотированных корпусов ХАНКО и НКРЯ (на примере создания картотек знаменательных и служебных частей речи).
3. Использование ИПС в составлении индивидуальной исследовательской картотеки.
4. Возможности использования материалов КОРГАЗ и American Rhetoric/ Online Speech Bank для выявления манипулятивных технологий СМИ.
5. Лингвистическое аннотирование зарубежных национальных мегакорпусов (на материале систем разметок корпусов изучаемых языков).
6. Прикладные возможности RusNet.
7. Выявление семантической наполненности смысложизненных категорий писателя (по конкордансным данным собрания «Весь Достоевский»).

8. Прикладные возможности устных зарубежных корпусов ANDOSL, BASE, CANCODE, COLT, CSLU Corpora, ELISA, EUSTAGE, FRED, IViE, L-CIE, LeaP, MICASE, Nationwide Speech Project Corpus, NECTE, SBCSAE.
9. Прикладные возможности программ CAT (с использованием текстовых оригиналов и переводческих версий, предложенных профессиональными переводчиками).
10. Параллельные подкорпусы СПИ.
11. Фольклорные корпусы и результаты применения их материалов в лингвофольклористике.
12. Проекты «Гуттенберг» и «The Oxford Text Archive».
13. Использование ресурсов НКРЯ и ХАНКО в ортологических целях.

Список самостоятельных работ:

1. Терминология корпусной лингвистики.
2. Связь корпусной лингвистики с другими теоретическими и прикладными областями науки.
3. Автоматическая обработка текста как проблема корпусной лингвистики.
4. Типология единиц языка (по данным частеречной разметки в корпусах флективных языков).
5. Коллекции речевых портретов подростков в корпусных собраниях.
6. «Семантические сети» как принцип устройства ментального лексикона и семантическая разметка как способ представления семантических групп в национальных корпусах.
7. Идиолекты как объект корпусной лингвистики.
8. Общая характеристика проектов WSC, TalkBank и Устного подкорпуса НКРЯ
9. Роль Roget's Thesaurus в выявлении семантических универсалий.
10. Параллельный подкорпус НКРЯ.
11. Жанровые коллекции «The Oxford Text Archive».
12. Complete Corpus of Old English и возможности его использования в диахронных исследованиях.
13. Ранговая статистика корпусных данных.

Список вопросов к зачету:

1. Объект и предмет корпусной лингвистики.
2. Понятие лингвистического компьютерного корпуса.
3. Понятие корпусной разметки.
4. Общая характеристика конкордансеров.
5. Национальные «мегакорпусы»: общая характеристика.
6. Технологические этапы создания лингвистически аннотированного корпуса.
7. Корпусная типология.

8. История создания компьютерных тезаурусов.
9. Идиолектный корпус (на выбор): общая характеристика.
10. Устный компьютерный корпус (на выбор): общая характеристика.
11. Переводческие программы.
12. Параллельный корпус (на выбор): общая характеристика.
13. Исторический корпус (на выбор): общая характеристика.
14. Прикладное применение корпусных материалов.

7. Учебно-методическое и информационное обеспечение дисциплины

а) основная литература:

Захаров В.П. Корпусная лингвистика. СПб.: Изд-во С.-Петерб. ун-та, 2005.

Национальный корпус русского языка и проблемы гуманитарного образования: [учебно-методическое пособие] / [отв. ред.-сост. Н. Р. Добрушина]. М.: ТЕИС, 2007.

б) дополнительная литература:

имеются в Научной библиотеке НГУ:

Зубов А.В. Информационные технологии в лингвистике : [учеб. пособие для вузов по спец. 021800 «Теорет. и прикл. лингвистика»] / А.В. Зубов, И.И. Зубова. М.: Академия, 2004.

Зубов А.В. Основы искусственного интеллекта для лингвистов : [учеб. пособие для вузов по спец. «Теорет. и прикл. лингвистика»] / А. В. Зубов, И. И. Зубова. М.: Логос, 2007.

российская:

Sinclair J. Preliminary Recommendations on Corpus Typology // http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Национальный корпус русского языка : 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.

Национальный корпус русского языка: 2003—2005. Сборник статей. М.: Индрик, 2005.

Азарова И.В. Санкт-Петербургский корпус агиографических текстов (СКАТ): формат XML-представления лингвистической информации и организации поиска данных на сайте / И. В. Азарова, Е. Л. Алексеева // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. Казань: Изд-во Казанского гос. ун-та, 2008. С. 3-6.

Белозёрова Н.Н., Чуфистова Л.Е. Шекспир и компания, или использование электронных библиотек при лингвистическом исследовании: Учеб. пособие. Тюмень: Издательство ТюмГУ, 2007.

Вадяев С.Е. Лингвистические принципы построения и использование корпуса текстов для исследования официально-делового стиля современного немецкого языка : На материале электронного корпуса «DER»: Дисс. ... канд. филол. наук. Самара, 2005.

Владимов Н.В. Корпусный подход к решению переводческих проблем: На материале письменных переводов с русского языка на английский: Дисс… канд. филол. наук. М., 2005.

Гвишиани Н.Б. Практикум по корпусной лингвистике = English on computer a tutorial in corpus linguistic : [учебное пособие по английскому языку]. М.: Высшая школа, 2008.

Ельмслев Л. Можно ли считать, что значения слов образуют структуру? / Л. Ельмслев; Пер. с англ. И.А. Мельчука. Синтаксические структуры / Н. Хомский; Пер. с англ. К.И. Бабицкого. Благовещенск : БГК им. А.И. Бодуэна де Куртенэ, 1998.

Коваль С. А. Лингвистические проблемы компьютерной морфологии. СПб.: Изд-во С.-Петерб. ун-та, 2005.

Коваль С.А. Роль корпуса в создании реалистичных моделей словоизменительной морфологии. URL: http://skowal.narod.ru/research/corpora2006/Koval_Corpora.2006.htm

«Корпусная лингвистика и лингвистические базы данных»: Материалы конф. СПб.: Изд-во С.-Петерб. ун-та, 2002.

«Корпусная лингвистика-2004»: Тез. докл.. СПб.: Изд-во С.-Петерб. ун-та, 2004 (а также труды участников конференций других лет).

Лагута О.Н., Лаврентьев А.М. Лингвистические приложения // Русские простонародные легенды и рассказы: сб. 1861 г. - Новосибирск: Наука, 2005. С. 185-317.

Лесников С.В. Автоматизированная лексикографическая система «Говор»: Дисс. ... канд. филол. наук. СПб., 1994.

Лесников С.В. Словарь русских словарей. М.: ГУП Моск. тип., 2002.

Малинина Ю.А. Сопоставительный анализ терминологии, сопровождающей Британский национальный корпус (British National Corpus) и Национальный корпус русского языка // Наука-2009: сборник научных статей. Ч. 2. Гродно, 2009. С. 121-123.

Марчук Ю.Н. Основы компьютерной лингвистики. - М.: Изд-во МПУ, 2000.

Орехов Б.В. Корпусная лингвистика и параллельный корпус переводов «Слова о полку Игореве» // Русский язык в поликультурной среде: лингводидактические и социокультурные проблемы высшего образования. Материалы Международной научно-практической конференции, посвященной Году русского языка и 450-летию добровольного присоединения Башкортостана к России 22–23 марта 2007 г. Уфа: УГНТУ, 2007. С.256-260.

Очерки научно-технической лексикографии: Материалы конф. «Корпус. лингвистика и лингвист. базы данных» / Под ред. А.С. Герда. СПб.: Изд-во С.-Петерб. ун-та, 2002.

Русская разговорная речь Заполярья. Норильск: Тексты: Материалы конференции «Корпусная лингвистика и лингвистические базы данных». СПб.: Изд-во С.-Петерб. ун-та, 2002.

Рыков В.В. Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru/c.html

Санцевич Н.А. Моделирование вариативности языковой картины мира на основе двуязычного корпуса публицистических текстов: Метафоры и семантические оппозиции : Автореф. дис. ... канд. филол. наук. М., 2004.

Токтонов А.Г. Новая лексика в русских газетах 1990-х годов: системно-словообразовательный анализ : На материале Компьютерного корпуса текстов русских газет конца XX века: Дисс… канд. филол. наук. М., 2006.

Чардин И.С. Лингвистические корпусы с разметкой на основе грамматики зависимостей и их применение при автоматическом синтаксическом анализе: Дисс… канд. филол. наук. М., 2004.

зарубежная:

Adolphs S. Introducing Electronic Text Analysis: A practical guide for language and literary studies. Abingdon: Routledge, 2006.

Aston G., Burnard, L. The BNC handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press, 1998.

Baker P. Using Corpora in Discourse Analysis. L.: Continuum, 2006.

Baker P., Hardie A., McEnery T. A Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press, 2006.

Biber D. University language: A corpus-based study of spoken and written registers. Amsterdam: John Benjamins, 2006.

Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating language structure and use. Cambridge: CUP, 1998. URL: http://books.google.com/books?id=2h5F7TXa6psC

Deignan A. Metaphor and Corpus Linguistics. Amsterdam: John Benjamins, 2005.

Developing linguistic corpora: a guide to good practice / Wynne M. (Ed.). URL: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

Gavioli L. aura. Exploring corpora for ESP learning. Amsterdam: John Benjamins, 2005.

Halliday M. A. K., Teubert W., Yallop C., Cermáková A. Perspectives in lexicology and corpus linguistics: An introduction. L.: Continuum, 2004.

Leech G. Corpus annotation schemes // Literary and Linguistic Computing. 1993. № 8/4. P. 275-281.

McEnery T., Wilson A. Corpus linguistics. Edinburgh : Edinburgh University Press, 2001.http://books.google.com/books?id=nwmgdvN_akAC

Meaningful texts: The extraction of semantic information from monolingual and multilingual Corpora / Barnbrook G., Danielsson P., Mahlberg M. (Eds.). L.: Continuum, 2005.

Meyer Ch. English corpus linguistics: An introduction. Cambridge: Cambridge University Press, 2002.

Studer P. Historical Corpus Stylistics: Media, Technology and Change. L.: Continuum, 2008.

Wallis S. Annotation, Retrieval and Experimentation or: you only get out what you put in // Annotating Variation and Change / A. Meurman-Solin, A.A. Nurmi (Ed.). Helsinki: Varieng, 2007 (э-версия на http://www.helsinki.fi/varieng/journal/volumes/01/wallis/)

в) программное обеспечение и Интернет-ресурсы:

Сайт Семинара по корпусной и компьютерной лингвистике ИЛИ РАН: http://corpora.iling.spb.ru/

David Lee's Corpus-based Linguistics Links: http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm

Rykov Corpus Linguistics home page: http://www.rykov-cl.narod.ru/

Лингвистические корпусы:

ALPINO Treebank http://odur.let.rug.nl/~vannoord/trees/

American Rhetoric/ Online Speech Bank http://www.americanrhetoric.com/

ANDOSL - Australian National Database of Spoken Language: http://andosl.anu.edu.au/andosl/

Arabic English Parallel News http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T18

ARCHER – A Representative Corpus of Historical English Registers: http://www.llc.manchester.ac.uk/research/projects/archer/

ARTFL Project (American and French Research on the Treasury of the French Language) http://artfl-project.uchicago.edu/ ,

AVIP – Archivio di Varietà di Italiano Parlato: ftp.cirass.unina.it

Bank of English (COBUILD - Collins Birmingham University International Language Database): http://www.titania.bham.ac.uk/docs/svenguide.html

BASE – British Academic Spoken English: http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/

BNC – British National Corpus http://www.natcorp.ox.ac.uk/

Brown Corpus http://khnt.aksis.uib.no/icame/manuals/brown/

CANCODE – Cambridge and Nottingham Corpus of Discourse in English: http://www.cambridge.org/elt/corpus/cancode.htm

CCA – Corpus of Contemporary Arabic, см. Latifa Al-Sulaiti homepage http://www.comp.leeds.ac.uk/eric/latifa/index.htm

CHILDES - Child Language Data Exchange System: http://childes.psy.cmu.edu/

COCA – Corpus of Contemporary American English http://www.americancorpus.org/

COLT - Bergen Corpus of London Teenage Language: http://torvald.aksis.uib.no/colt/

Complete Corpus of Old English http://www.doe.utoronto.ca/

Corpus del Español http://www.corpusdelespanol.org/

CSLU Corpora – корпус Center for Spoken Language Understanding: http://www.cslu.ogi.edu/corpora/

DCPSE - Diachronic Corpus of Present-Day Spoken English (о проекте: http://www.ucl.ac.uk/english-usage/projects/dcpse/)

EEKK- Estonian Emotional Speech Corpus (о проекте: http://193.40.113.40:5000/docs/readme.html)

ELISA – тюбингенский English Language Interview Corpus as a Second-Language Application: http://www.uni-tuebingen.de/elisa/html/elisa_index.html

Emotional Speech Corpus (о проекте: http://arrow.dit.ie/cgi/viewcontent.cgi?article=1004&context=dmcart)

EUSTAGE – Edinburgh University Speech Timing Archive and Corpus of English: http://www.cstr.ed.ac.uk/projects/eustace/

Finnish Emotional Speech Database (о проекте: http://www.mediateam.oulu.fi/publications/pdf/394.pdf)

FRED – Freiburg English Dialect Corpus: http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/

ICLE - International Corpus of Learner English http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm

IViE - Intonational Variation in English: http://www.phon.ox.ac.uk/files/apps/IViE//

Lancaster Corpus of Children's Project Writing http://www.lancs.ac.uk/fass/projects/lever/index.htm

L-CIE - Limerick Corpus of Irish-English: http://www.ul.ie/~lcie/

LeaP - Learning the Prosody of a foreign language: http://www.phonetik.uni-freiburg.de/leap/

Longman Spoken American Corpus (о проекте: http://www.pearsonlongman.com/dictionaries/pdfs/Spoken-American.pdf)

MICASE - Michigan Corpus of Academic Spoken English http://micase.elicorpora.info/

Nationwide Speech Project Corpus: http://www.ling.ohio-state.edu/~cclopper/nsp/index.html

NECTE - Newcastle Electronic Corpus of Tyneside English: http://research.ncl.ac.uk/necte/

Northern Ireland Transcribed Corpus (о нем см. стать. J. Kirk в сборнике «New directions in English language corpora: methodology, results, software» http://books.google.ru/books?id=iBf33Lu--3kC&pg=PA65&lpg=PA65&dq=Northern+Ireland+Transcribed+Corpus&source=bl&ots=rat4EQOgzC&sig=5bKFFfexCtQPoo3qwrbsXI_88eA&hl=ru&ei=wu5rS8WJFY6UmAOxr_T5BA&sa=X&oi=book_result&ct=result&resnum=1&ved=0CAcQ6AEwAA#v=onepage&q=Northern%20Ireland%20Transcribed%20Corpus&f=false)

Roget's Thesaurus of English Words and Phrases http://www.gutenberg.org/files/10681/10681-body.txt

SBCSAE - Santa Barbara Corpus of Spoken American English: http://projects.ldc.upenn.edu/SBCSAE/

Spoken Corpus of the Survey of English Dialects (о проекте: http://www2.iath.virginia.edu/ach-allc.99/proceedings/scott.html)

TalkBank: http://talkbank.org/

The Oxford Text Archive http://ota.ahds.ac.uk/

WSC – Wellington Corpus of Spoken New Zealand English http://khnt.hit.uib.no/icame/manuals/wsc/INDEX.HTM

YCOE - The York-Toronto-Helsinki Parsed Corpus of Old English Prose: http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

Весь Достоевский http://petrsu.ru/~Dostoevsky/main.htm

ГАФИС «Брюсов» www.vbryusov.ru

КОРГАЗ – Компьютерный корпус текстов русских газет конца XX века http://www.philol.msu.ru/~lex/corpus/

Корпус китайского языка http://ccl.pku.edu.cn/YuLiao_Contents.Asp

Корпус русского языка XVIII века (о проекте: http://textualheritage.org/content/view/206/93/)

Корпус старолатышских текстов http://www.korpuss.lv/senie/toc.jsp

Лидские корпусы русских текстов corpus.leeds.ac.uk(

НКРЯ – Национальный корпус русского языка ruscorpora.ru

СКАТ http://project.phil.pu.ru/scat/page.php?page=project

Словарь-конкорданс публицистики Ф.М. Достоевского http://dostoevskii.karelia.ru/index.phtml

Словарь-корпус языка А. С. Грибоедова http://feb-web.ru/feb/concord/abc/11/gr11-522.htm и http://www.inforeg.ru/electron/concord/concord.htm

СПИ - Историко-параллельный корпус «Слова о полку Игореве» http://nevmenandr.net/slovo/

Упсальский корпус современных русских текстов http://www.slaviska.uu.se/korpus.htm

ХАНКО – Хельсинкский аннотированный корпус русских текстов http://www.helsinki.fi/venaja/russian/e-material/hanco/index.htm

8. Материально-техническое обеспечение дисциплины

На лекциях и при выполнении практических и самостоятельных работ необходимы компьютеры с доступом в Интернет.