Министерство образования и науки РФ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
(ФГБОУ ВО «АмГУ»)

КОРПУСНАЯ ЛИНГВИСТИКА

сборник учебно-методических материалов
для направления подготовки 45.04.03 – Фундаментальная
и прикладная лингвистика

Благовещенск 2017

Печатается по решению
редакционно-издательского совета
филологического факультета
Амурского государственного
университета

Составитель: Морозова О.Н.

Корпусная лингвистика: сб. учеб.-метод. материалов для направления подготовки 45.04.03 «Фундаментальная и прикладная лингвистика. – Благовещенск : Изд-во Амур. гос. ун-та, 2017. http://irbis.amursu.ru/DigitalLibrary/AmurSU_Edition/8303.pdf

Содержание

МАТЕРИАЛЫ К ЛЕКЦИЯМ

Корпусная лингвистика В.В. Рыкова.

Сопоставление корпусной и традиционной лингвистик, а также традиционного лингвиста и "корпусного". Очень условное и преувеличенное - для понимания сущности описываемого подхода.

Корпусная лингвистика <==> Традиционная лингвистика

1 Основное внимание – изучение речи <==> Основное внимание – изучение языка

2 Цель –описание языка в том виде, как он проявил себя в речи, представленной в виде специально подобранного корпуса текстов <==> Цель – описание и объяснение языка

3 В своих исследованиях опирается на данные корпуса текста <==> В своих исследованиях идёт от теории к её объяснению и подтверждению в фактах речи

4 Предпочитает квантитативные методы <==> Предпочитает квалитативные методы

5 Видит себя частью традиций, базирующихся на эмпирических методах <==> Видит себя частью традиций, базирующихся на рационалистических методах

6 Текст рассматривается как некоторая физическая сущность <==> Текст рассматривается как некоторая абстракция

7 Составление грамматики конкретных языков <==> Изучает языковые универсалии

8 Основное внимание уделяется форме <==> Основное внимание – не только форме, но и содержанию

9 Рассматривает тексты в глобальной перспективе <==> Рассматривает тексты в локальной перспективе

10 Фокусирует своё внимание на как можно более широком взгляде на текст, неограниченном ни какими догмами <==> Анализирует некоторую конкретную, искусственно ограниченную, проблемную область

11 В своих выводах опирается на наблюдение речевой деятельности, проявленной в виде текстов <==> Опирается на интуицию в отборе речевого материала, в отборе эмпирических материалов своих исследований

12 Часто пользуется вероятностными методами и статистикой для первичной обработки речевого материала <==> Предпочитает логические рассуждения

13 Проводится работа с лингвистическими данными (словоупотреблениями) в том виде, в каком они встречались в контексте <==> Предпочитаются искусственные примеры, из изолированных от текста словоупотреблений

14 Предпочитает индуктивные методы обработки эмпирического словесного материала, считает их сутью научного метода <==> Предпочитает дедуктивные методы обработки эмпирического словесного материала

15 Верит в научные открытия, основанные на обработке эмпирических данных <==> Верит в открытия, основанные на процедурах, оценках, сравнениях и т.д., т.е., как результат многовековых исследований

Стадии работы:

Корпусная лингвистика имеет дело с уже собранным материалом.

1.Необходимо представить структуру речевой действительности.

2.Выявить, какие материальные ограничения есть на составление корпуса.

3.Отбор текстов и составление корпуса текста.

4.Компиляция корпуса.

Определение корпуса текстов:

1.Некоторое собрание текстов.

2.В основе лежит логический замысел, логическая идея, объединяющая эти тексты.

3.Воплощение логической идеи:

правила организации текстов в корпус

алгоритмы и программы анализа корпуса текстов
сопряжённая с этим идеология и методология
4.Корпус текстов принадлежит к четвёртой фактуре речи.

*Фактуры речи:
устная речь
письменная речь
печатная речь
тексты на машинном носителе

Корпус текстов:
некоторый филологический объект;
организованное словесное множество, элементами которого являются определённым образом отобранные тексты;
организованное определённым образом словесное единство, элементами которого являются тексты или специальным образом отобранные отрывки из текстов.

Самые популярные, распространенные, важные корпуса текстов.

	Название	Год	Количество словоупотреблений	Язык
1	BUC	1964	1 000 000	Англ.(USA) печатный
2	AHI	1971	5 000 000	Англ.(USA) печатный
3	LOB	1978	1 000 000	Англ.(G.B) печатный
4	Birmingem corpus	1987	20 000 000	Англ.(G.B) печатный
5	Kolhapur corpus	1988	1 000 000	Англ.(Индийский)
6	TOSCA	1988	1.5 000 000	Англ.(G.B) печатный
7	SEU Corpus	1989	1 000 000	Англ.(G.B) печатный
8	CHILDES	1990	20 000 000	Англ.(детский) устный
9	Nijmengen	1991	132 000 000	Англ.(G.B) печатный, устный
10	LLELC	1991	50 000 000	Англ. печатный, устный
11	Map Task Corpus	1991	147 000 000	Англ.(Scotland) устный
12	LCLE	1992	10 000 000	Англ. печатный (для иностранцев)
13	SEC	1992	53 000 000	Англ.(G.B) устный
14	Wellington Corpus	1993	1 000 000	Англ.(Новозеланд.) печатный
15	POW	1993	65 000 000	Англ.(детский) устный
16	BNC	1995	100 000 000	Англ.(G.B) устный, письменный, печатный
17	Corpus of Spoken	1991	2 000 000	Англ.(USA) устный
18	ICLE	1997	200 000 000	Англ. письменный (для иностранцев)
19	Bank of English	1997	320 000 000	Англ.(G.B) печатный

Основная задача компьютерной лингвистики.
Полное и системное отражение содержательного общения на языке. Основной особенностью направления исследования, которое можно назвать информационно-семиотическое направление лингвистических исследований, является подход к рассмотрению прикладных проблем лингвистики строго в коммуникативных процессах. При этом в центре внимания оказывается не язык (естественный), как система, и не проблема его формализации (имеющая самые различные толкования), а процесс содержательного общения на языке, и по возможности точное его описание, которое может быть использовано для решения научно-технических задач информатики.
Первая попытка достаточно полного и системного отражения “содержательного общения на языке” в сфере печатной речи была предпринята составителями BUC.

Корпусная лингвистика сделала возможным:
1. Уточнить результаты и выводы, проведённых ранее исследований речи.
2. Произвести новые, более широкие и системные по охвату эмпирического речевого материала лингвистические исследования..

В центре внимания корпусной лингвистики оказалась языковая личность, т.е., её речевая деятельность, массовая коммуникация, проблема её описания.

Главная цель:
лингвистическое описание языковой системы (подход от конкретного изучения коммуникации людей),
особый способ отражения речевого материала в корпусе текстов, который может использоваться в свою очередь другими лингвистическими дисциплинами.

Корпусная лингвистика имеет две черты, дающие основание претендовать на положение самостоятельной дисциплины:
1. Характер используемого словесного материала.
2. Специфика инструментария.

Таким образом, корпус текстов, с одной стороны, это исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, результат деятельности корпусной лингвистики.

"Отступления" корпусной лингвистики:
1. КЛ не отрицает ценности и необходимости речевых данных не представленных в корпусной форме;

КЛ утверждает то, что из корпуса текстов невозможно извлечь все возможные лингвистические выводы, т.е., что корпус текстов не является самодостаточным.

Классификация корпуса текстов.
По степени организации и структурированности
1. Электронный архив – это тексты на электронном носителе, но их форма представленая на машинном носителе не стандартизирована и не унифицирована.
2. Электронная библиотека – тексты здесь представлены однородным и стандартизированным образом.
3. Корпус текстов – форма стандартизирована и унифицирована, тексты предназначены для отражения части лингвистической реальности.
4. Субкорпус – это некоторая автономная часть корпуса.

По хронологическому признаку:
1. Синхронический;
2. Мониторный (отслеживает текущее состояние языка);
3. Диахронический.

По индексации:
1. Простой;
2. Аннотированный.

По языку:
1. Одноязычный;
2. Двуязычный;
3. Многоязычный.

По способу применения и использования корпуса:

1. Исследовательский;
2. Иллюстративный;
3. Параллельный.

По способу существования корпуса:
1. Динамический;
2. Статический.

Программы анализа корпуса.
1. Программы составления конкордансов.
2. Программы индексирования или аннотирования.

Конкорданс – список словоформ встречающихся в тексте, расположенных в алфавитном порядке. В противоположногсть словарю – слово даётся с его словесным окружением.

Конструирование и применение корпусов.
Единой методики для всех языков нет. Так как разные языки, традиции, технологические процессы. Но основные требования таковы:
1. Кто пользователь корпуса? (индивид, группа, лингвистическое общество).
2. Какова логическая идея, которая положена в основу корпуса?
3. С каким объёмом данных мы будем работать при составлении корпуса? На сколько это необходимо и реалистично?
4. Используем отрывки из текстов, полные тексты или то и другое.
5. Процедура отбора текстов в корпусе. Для разных целей по-разному:
обследование речевого материала,
сканирование текстов,
окончательное формирование, составление корпуса.
6. Стандартизированное представление корпуса на уровне отраслевых стандартов, т.е., представление всего корпуса как продукта:
аннотация всего текста в целом
унифицированное представление словесного материала текста.
7. Аннотирование, индексирование словесного материала текста.

Лингвистические исследования, базирующееся на корпусе текстов.
Применение корпусов текста в исследовании языка.
1. Подбор нужного корпуса текстов:
доступность,
достаточность словесного материала,
является ли данный корпус представительным для поставленной задачи,
каким образом были отобраны тексты,
достоверно ли представление индексов (если он индексирован).
2. Насколько необходимо данное исследование (адоптация целей и задач исследования под наличный корпус текстов).
3. Практические рекомендации:
анализировать то, что ясно и явно представлено в машиной форме,
искать то, что легко найти,
подсчитывать то, что легко подсчитывается.

Проблемная область.
Это область реализации языковой системы, содержащая феномены, подлежащие лингвистическому описанию. Проблемная область для конкретного корпуса может быть сколь угодно велика или сколь угодно мала. Всё определяется выбранным объектом анализа.

В идеале проблемная область имеет 2 измерения: 1) языковое измерение, проявляющееся в существовании потенциальной возможности, появления других употреблений, дополняющих массив имеющихся реализаций; 2) речевое измерение, представленное речевыми высказываниями.

В корпусной лингвистике, как правило, языковой аспект фактически игнорируется, т.к. изначально фиксируется область привлекаемых языковых данных – реализации языковой системы. Однако для регулярно изменяемых корпусов данных, языковой аспект проблемной области сразу проявляется при разработке принципов модификации корпуса. Кроме того, для лингвистического исследования (кроме специально оговариваемых случаев) в центре внимания стоит именно языковое измерение, т.к. его следует реконструировать в результате анализа.

Репрезентативность.
Важнейшее свойство корпуса текстов – его репрезентативность. Т.е., способность отражать все свойства проблемной области. Соблюдаются ли пропорции, которые наблюдаются проблемной области. Простейший способ преобразования проблемной области в корпус это пропорциональное сужение проблемной области. Репрезентативность определяется параметрами: фонетическими, морфологическими, синтаксическими, стилевыми.

Корпусная и компьютерная лингвистики
Под корпусной лингвистикой понимается раздел лингвистики, занимающийся разработкой и использованием лингвистических корпусов данных. Прежде, чем говорить о собственно корпусной лингвистике как о научной дисциплине, необходимо определить понятие *корпуса*. Как такового общепринятого определения пока выделено не было, поэтому приведём несколько наиболее популярных:

- корпус — это организованное определённым образом словесное единство, элементами которого являются тексты или специальным образом отобранные отрывки из текстов
- корпус – это набор лингвистических данных из определённого языка в форме записанных высказываний или письменных текстов, доступный для анализа
- корпус — это набор естественных текстов на любом языке, устных или письменных, который хранится в электронном виде и позволяет организовать компьютеризированный поиск

В целом, **корпус данных** представляет собой сформированную по определенным правилам выборку данных из т.н. **проблемной области**, т.е. по сути, корпус данных представляет собой результат отображения проблемной области. Под проблемной областью понимается область реализаций языковой системы, содержащая феномены, подлежащие лингвистическому описанию. Проблемная область для конкретного корпуса данных может быть сколь угодно велика или мала — все определяется выбранным объектом анализа. В идеале, проблемная область имеет два измерения — языковое и речевое. **Речевое измерение** представлено речевыми высказываниями, или реализациями, а **языковое измерение** проявляется в существовании потенциальной возможности появления *других* употреблений, дополняющих массив имеющихся реализаций. Как правило, корпусная лингвистика практически полностью игнорирует языковой аспект, поскольку изначально в исследованиях в рамках этой дисциплины фиксируются именно реализации языковой системы. Такой подход обусловлен тем, что зафиксировать возможно только реально существующие единицы, а не «потенциальной возможности» их появления. Однако для регулярно изменяемых корпусов данных языковой аспект проблемной области дает о себе знать на стадии разработки принципов модификации корпуса. Кроме того, для лингвистического исследования в целом (кроме специально оговариваемых

случаев) в центре внимания стоит именно языковое измерение, поскольку именно его следует реконструировать в результате анализа. **С чисто практической точки** зрения проблемную область можно определить как множество данных, обработка которых затруднена из-за того, что языковых реализаций слишком много.

В отличие от проблемной области, **корпус данных** имеет только одно измерение — речевое, поскольку сам по себе он не обладает возможностью производства своих составляющих. Это, однако, не означает, что корпус данных не может использоваться для реконструкции языка как системы. Напротив — это одна из основных задач лингвистического исследования корпуса. Выводы о функционировании языка как системы делаются исследователями-лингвистами на основе отдельных результатов деятельности языка.

Отдельного обсуждения заслуживает проблема выделения единиц хранения корпуса данных. **Единица хранения** — это некоторая совокупность естественно-языковых выражений проблемной области, которой сопоставляется одно описание на некотором метаязыке, определяемом процедурой формирования корпуса. Поскольку корпус данных представляет собой выборку из проблемной области, сформированную по некоторым определенным принципам, единица хранения непосредственно зависит от оснований, по которым осуществлялась выборка. В зависимости от этих оснований и от цели исследования, единицами хранения корпуса могут быть отдельные слова, короткие фразы, предложения, словосочетания (синтагмы). Если корпус предполагается для синтаксического анализа, то он должен включать целые тексты или достаточно большие их фрагменты. На основании описания единицы хранения можно судить о том, какая часть проблемной области представлена в корпусе. Например, единица хранения корпуса рекламных слоганов, созданного в Отделе экспериментальной лексикографии Института русского языка РАН, включает следующие характеристики: слоган: *Для мужчин, которые любят женщин, которые любят мужчин*; фирма: *«Louis Azzaro»*; предмет: *туалетная вода Azzaro pourHomme*; область: *косметика и парфюмерия*; вид слогана: *перевод с французского*; оригинал: Pour les hommes qui aiment les femmes qui aiment les homes; источник: *Космополитен.* Таким образом, выражение естественного языка «*Для мужчин, которые любят женщин, которые любят мужчин*» и сопоставленные ему характеристики вместе образуют единицу хранения, которая может вводиться в базу данных или включаться в обычный файл текстового формата.

Виды корпусов данных: **Исследовательский корпус** – т.е. корпус, который предназначен для изучения различных аспектов функционирования языковой системы. Такие корпуса строятся не post factum - т.е. после проведения какого-либо исследования, а до его проведения. **Иллюстративный корпус** – т.е. корпус, который создается после проведения научного исследования: целью здесь является не столько выявить новые факты, сколько подтвердить и обосновать уже полученные результаты. Такие корпусы не являются статистически правильным отображением проблемной области, т.к. они включают лишь то, что достаточно для иллюстрации описываемого феномена. **Статический корпус** – корпус, отражающий определенное временное состояние языковой системы. Типичными представителями этого вида корпусов являются авторские корпусы — т.е. коллекции текстов писателей. **Динамический (мониторный) корпус** – отличается от статического тем, что не предполагает раз и навсегда заданного набора текстов. В течение заранее фиксированного промежутка времени происходит обновление и/или дополнение множества текстов корпуса с целью мониторинга состояния проблемной области и динамики её изменения. Специфика эксплуатации динамического корпуса состоит в том, что пользователь при проведении исследования может выделить из общего генерального корпуса рабочий корпус, включающий лишь часть текстов генерального корпуса.

Важнейшие свойства корпуса: **репрезентативность** по отношению к проблемной области – т.е. способность корпуса текстов отражать все релевантные для данного исследования свойства проблемной области в определенной пропорции, которая определяется частотой

встречаемости данного явления в проблемной области. Т.е., частота появления некоторого явления в лингвистическом корпусе должна быть близка частоте появления этого явления в соответствующей проблемной области; **полнота** – требует учета релевантных явлений в корпусе, даже если это не соответствует идее пропорционального сужения между корпусом и проблемной областью. Требование полноты совершенно необходимо в тех случаях, когда лингвист-конструктор корпуса лишь приблизительно знает, что ему необходимо искать. В такой ситуации исследовательский корпус может приобрести те или иные черты иллюстративного корпуса; **экономность** – корпус текстов должен экономить усилия исследователя при изучении проблемной области. В частности, он должен быть не просто строгим подмножеством текстов проблемной области, но, по возможности, существенно отличаться от нее по объему (с сторону уменьшения). Корпус считается тем более «экономичным», чем выше порог отображения явлений; **структурированность** хранимого материала.

Основные задачи корпусной лингвистики: **первичной задачей** корпусной лингвистики считается объективное лингвистическое описание языковой системы, причём к этому описанию корпусная лингвистика подходит, отталкиваясь от изучения конкретной человеческой коммуникации, от реальных текстов. В качестве **вторичной задачи** рассматривается выработка особого способа отражения речевого материала в корпусе текстов. Этот способ, в свою очередь, может использоваться другими лингвистическими дисциплинами. Ещё одна часто выделяемая задача корпусной лингвистики заключается в изучении *вероятности* лингвистических явлений (в отличие от традиционной лингвистики, которая изучает их (явлений) *возможность*; Так, например, традиционная лингвистика скажет, что конструкция I'm not в литературном английском возможна, а конструкция I ain't – нет. Корпусная же лингвистика не скажет, что конструкция «I ain't» невозможна – она скажет, что эта конструкция маловероятна).

Основные направления научной деятельности в рамках корпусной лингвистики: **Во-первых**, это лексикографические исследования, создание словарей. Практически все современные словари английского языка (Collins, Webster, MacMillan и т.д.) издаются на основе огромных корпусов, которые позволяют сделать словарь репрезентативным. То есть, словарь может быть верным или не верным относительно данного корпуса. **Во-вторых**, изучение корпусов позволяет получать точные данные о лексическом составе языков, об относительных частотах употребления тех или иных слов. В частности, при помощи корпусной лингвистики был окончательно доказан так называемый закон Ципфа, утверждающий, что если в любом естественном языке все слова упорядочить по убыванию частоты их использования, то частота любого слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру (так называемому рангу этого слова). Например второе по частоте слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и так далее. **В-третьих**, корпусная лингвистика изучает и изменения в лексическом составе языков, различные его вариации (например, появление и исчезновение неологизмов). **В-четвертых,** корпусная лингвистика изучает грамматики естественных языков, в частности – сочетаемости тех или иных грамматических явлений друг с другом. Естественно, что данные, полученные из живой речи, гораздо более актуальны, чем умозрительные грамматики традиционной лингвистики. **В-пятых**, корпусная лингвистика занимается изучением текстов. Например, используя корпусы текстов, можно научиться определять функциональный стиль текста через его статистические характеристики – среднюю длину слова и предложения, характерные сочетания слов и т.д. Такие методы уже существуют и используются в автоматическом реферировании и тематическом поиске. **В-шестых**, корпусная лингвистика активно используется в лингводидактике, то есть, в обучении иностранным языкам. Чтобы знать, чему, учить, необходимы точные количественные данные о преподаваемом языке — состав наиболее частотной лексики, вероятности употребления тех или иных грамматических конструкций и т. д. **В-седьмых**, корпусная лингвистика

занимается проблемами машинного перевода, для чего строятся и используются т.н. многоязычные выровненные (параллельные) корпусы, в которых каждой фразе на одном языке сопоставлен её эквивалент на другом языке. Кроме машинного перевода, такой корпус можно использовать для исследований, связанных со сравнением оригинальных и переводных текстов.

Соотношение корпусной и компьютерной лингвистики. Как уже упоминалось выше, компьютерной лингвистикой называется ветвь лингвистики, занимающаяся моделированием языка с использованием компьютерной техники. Корпусная лингвистика занимается примерно тем же, так что можно сказать, что эти дисциплины дополняют друг друга. Компьютерная лингвистика, например, создаёт инструменты (то есть, программы) для корпусной лингвистики. Например, исследователям в области корпусной лингвистики необходимы средства для автоматической разметки классов слов в корпусах. А если имеется корпус на, скажем 100 миллионов словоупотреблений и необходимо отметить часть речи у каждого слова, то вручную это сделать совершенно нереально. Тут и понадобится специализированное программное обеспечение. Кроме того, очень активно в современном мире используются программы морфологического и синтаксического анализа. Их обучение также происходит на основе корпусов текстов. Кроме того, для исследования корпуса бывает важно сначала снять лексическую неоднозначность, то есть, выделить слова-омонимы (лук, кисть). В большом корпусе сделать это вручную затруднительно, поэтому компьютерная лингвистика создаёт программы семантического анализа текстов, которые способны в автоматическом режиме определять, в каком значении употреблено то или иное слово. И, наконец, компьютерная лингвистика активно занимается вопросами создания параллельных корпусов, о которых говорилось выше.

Корпусная лингвистика А.Б. Кутузова

Понятие лингвистического корпуса Прежде, чем говорить о корпусной лингвистике, необходимо определить само понятие лингвистического корпуса. По-английски это будет linguistic corpus или text corpus, множественное число linguistic corpora (corpuses употребляется реже). Существует довольно много определений, которые сходятся в одном: корпус есть «некоторый филологический объект». Вот несколько дефиниций: ● корпус — это организованное определённым образом словесное единство, элементами которого являются тексты или специальным образом отобранные отрывки из текстов; ● корпус – это набор лингвистических данных из определённого языка в форме записанных высказываний или письменных текстов, доступный для анализа; ● корпус — это набор естественных текстов на любом языке, устных или письменных, который хранится в электронном виде и позволяет организовать компьютеризированный поиск; ● пожалуй, наиболее полное определение: корпус есть собрание отрывков текстов в электронной форме, отобранных в соответствии с внешними критериями, чтобы наиболее полно представлять язык или вариацию языка. Функционирует как источник данных для лингвистических исследований. (John Sinclair)

Вот примеры корпусов: ● тексты конкретного писателя или писателей; ● тексты за конкретное десятилетие или столетие; ● современные тексты определённой тематики; ● современные тексты, адекватно представляющие язык или общество. В одном из определений было сказано, что корпус может быть как устным, так и письменным. Вообще, существует мнение, что лингвистические корпусы не являются ни устными, ни письменными, ни печатными, а представляют собой четвёртую фактуру речи — тексты на машинном носителе — тот самый digital text. Впрочем, с этим взглядом можно спорить. Понятно, что корпус — это набор текстов, с которыми можно что-то делать. Но что же может делать корпус? Ответ может показаться неожиданным: сам корпус не может делать ничего. Но мы можем использовать специальное программное обеспечение, чтобы искать в корпусе что-либо и производить некоторые вычисления. Что же мы можем искать? В

первую очередь, это слова и фразы, которые имеют культурную или лингвистическую значимость.

Кроме того, предметом поиска могут являться какие-либо пометки, которые вы добавили к корпусу, например, пометка «существительное». А вот примеры того, что может нам выдать поиск по корпусу: ● все употребления выбранного слова в непосредственном контексте; ● вариации и последовательность в использовании лексики; ● слова, которые чаще всего стоят рядом с выбранным словом; ● наиболее важные различия между двумя наборами текстов; ● как тот или иной писатель использует слова и фразы; ● интертекстуальность: значение слова как сумма его употреблений; ● скрытые (потенциальные) модели использования лексики; ● развитие концептов во времени; ● сравнение языков. В частности, нам, как переводчикам, наиболее актуальны возможности поиска контекстов слов, имеющих несколько переводных эквивалентов, а также подбор эквивалентов терминологических и фразеологических словосочетаний в параллельных корпусах, о которых мы будем говорить в следующих лекциях. Важнейшее свойство корпуса – репрезентативность, то есть, способность отражать все свойства проблемной области. Репрезентативность определяется фонетическими, морфологическими, синтаксическими и стилевыми параметрами корпуса. Именно репрезентативность отличает корпус от простого набора текстов. Не в последнюю очередь репрезентативность зависит от размера корпуса.

Эмпирический подход в сравнении с хомскианской лингвистикой Некоторые русскоязычные источники указывают, что впервые идея о том, что достоверные лингвистические данные могут быть получены лишь из большого массива текстов, была высказана Р.Г. Пиотровским в 60-х годах. На самом деле, осмысленные исследования в области корпусов начались ещё в сороковые годы (Блумфилд, Фрайс и Бонджерс). Но в 50-60-е годы возобладала концепция Ноама Хомского1 (хомскианская лингвистика, chomskyan linguistics). Она заключалась в том, что нужно изучать лишь competence (языковое знание, «язык» по Соссюру), а не performance (языковое употребление, «речь» по Соссюру). Ведь число высказываний естественного языка бесконечно, поэтому исследовать их бессмысленно. С другой стороны, количество языковых правил, которые и составляют competence, конечно. Поэтому их можно исследовать. Таким образом, произошёл уход от эмпирики в сторону рационализма и интроспекции (использования интуиции носителей языка). Тем не менее, некоторые учёные продолжали использовать корпусные методики и в период безраздельного господства генеративной лингвистики. Причина повышения интереса к корпусным исследованиям в последнее время — появление компьютеров, которые сделали возможной обработку огромных массивов текстов. Кроме того, всё больше учёных склоняется к тому, что интроспекция как метод изучения языка не всегда адекватна, и более научно опираться на естественные данные. Известные корпусные лингвисты Тони Мак-Эннери и Эндрю Уилсон пишут, что нужно использовать и эмпирику, и интроспекцию, и искусственные данные, и естественные. Корпусная лингвистика ни в коем случае не отрицает ценности и необходимости речевых данных, не представленных в корпусной форме. Кроме того, из корпуса текстов невозможно извлечь все возможные лингвистические выводы, то есть, корпус текстов не является самодостаточным.

Так, Чейф считает, что корпусный лингвист должен не только описывать явления языка, но и стараться объяснить их. Вообще, в центре внимания корпусной лингвистики оказалась языковая личность, то есть, её речевая деятельность, массовая коммуникация, проблема её описания. В этой таблице (её автор — Владимир В. Рыков) показаны основные отличия корпусной лингвистики от традиционной (хомскианской):

Корпусная лингвистика	*Традиционная лингвистика*
Основное внимание – изучение **речи**	Основное внимание – изучение **языка**
Цель – описание языка в том виде, как он проявил себя в речи, представленной в виде специально подобранного корпуса текстов	Цель – описание и объяснение языка
В своих исследованиях опирается на данные корпуса текста	В своих исследованиях идёт от теории к её объяснению и подтверждению в фактах речи
Предпочитает **квантитативные (количественные)** методы	Предпочитает **квалитативные (качественные)** методы
Видит себя частью традиций, базирующихся на **эмпирических** методах	Видит себя частью традиций, базирующихся на **рационалистических** методах
Текст рассматривается как некоторая **физическая сущность**	Текст рассматривается как некоторая **абстракция**
Составление грамматики **конкретных** языков	Изучает языковые **универсалии**
Основное внимание уделяется **форме**	Основное внимание – не только форме, но и **содержанию**
Рассматривает тексты в **глобальной** перспективе	Рассматривает тексты в **локальной** перспективе
Фокусирует своё внимание на как можно более **широком** взгляде на текст, неограниченном ни какими догмами	Анализирует некоторую **конкретную**, искусственно ограниченную, проблемную область
В своих выводах опирается на наблюдение речевой деятельности, проявленной в виде текстов	Опирается на интуицию в отборе речевого материала, в отборе эмпирических материалов своих исследований
Часто пользуется **вероятностными методами** и **статистикой** для первичной обработки речевого материала	Предпочитает **логические рассуждения**
Проводится работа с лингвистическими данными (словоупотреблениями) в том виде, в каком они встречались в контексте	Предпочитаются искусственные примеры, из изолированных от текста словоупотреблений
Предпочитает **индуктивные** методы обработки эмпирического словесного материала, считает их сутью научного метода	Предпочитает **дедуктивные** методы обработки эмпирического словесного материала
Верит в научные открытия, основанные на **обработке эмпирических данных**	Верит в открытия, основанные на **процедурах, оценках, сравнениях** и т.д.

История корпусной лингвистики Собственно, корпусы люди составляли и изучали ещё до появления корпусной лингвистики, начиная с XVIII века. Примеры: исследования Библии (Cruden и многие другие), составление словарей (Johnson, Oxford English Dictionary, Webster Dictionary), преподавание языков (частотный корпус Thorndike'a, 1921), дескриптивная грамматика (Fries, 1940, Quirk, 1968). Корпус Квирка (Survey of English Usage) включал один миллион словоупотреблений и изначально представлял собой один миллион карточек размером 6 на 4 дюйма, 17 строк текста на каждой. Этот корпус стал последним не электронным. Его составление заняло 25 лет, и к 1989 году, когда он был закончен, технология ушла далеко вперёд. Пришлось срочно переводить корпус в цифровую форму. Теперь этот корпус доступен в Юниверсити Колледж в Лондоне. Основные вехи создания компьютерных корпусов: 1. 1960-е: Брауновский корпус, (США), 1 млн. слов 2. 1970-е: LOB корпус (Великобритания, Норвегия), 1 млн. слов 3. 1980-е: Машинный Фонд русского языка 4. Уппсальский корпус русского языка (Швеция), 1 млн. слов 5. 1990-е: British National Corpus, 100 млн. слов, национальные корпуса (венгерский, итальянский, хорватский, чешский, японский) объёмом 100 млн. слов 6. The Bank of

English, Birmingham (Collins Cobuild), 600 млн. слов 7. 2000-е: American National Corpus, 100 млн. слов 8. Corpus of Contemporary American English, 400 млн. слов. 9. Национальный корпус русского языка, 140 млн. слов 10. Gigaword corpora: английский, арабский, китайский, 2 млрд. слов 11. Oxford English corpus, 2 млрд. слов. Таковы основные продукты деятельности корпусной лингвистики на сегодняшний день. В.В. Рыков даже пишет, что корпусная лингвистика – спорный термин, так как непонятно, имеется ли в виду наука о том, как создавать корпусы или же лингвистика, основанная на данных из корпусов. На практике, обычно под корпусной лингвистикой понимают и то, и другое. То есть, корпус для корпусной лингвистики, с одной стороны, исходный речевой материал, с другой – результат деятельности.
Подытоживая: Корпусная лингвистика сделала возможным: 1. Уточнить результаты и выводы проведённых ранее исследований речи. 2. Произвести новые, более широкие и системные (по охвату эмпирического речевого материала) лингвистические исследования

Основные задачи и направления корпусной лингвистики. Взаимодействие корпусной лингвистики и компьютерной (computational) лингвистики

Как уже говорилось в предыдущей лекции, деятельность в рамках корпусной лингвистики может быть сведена к созданию корпусов и к лингвистическим исследованиям на их базе (все задачи по изучению больших массивов текстов). В какомто смысле, корпусная лингвистика сама создаёт свой материал, точнее, самостоятельно структурирует его. Именно это делает её самостоятельной лингвистической дисциплиной – у неё специфический характер используемого словесного материала (корпусы) и свой собственный инструментарий (программы анализа корпусов). А самостоятельность науки как раз и определяется наличием у неё собственного материала, либо собственных методов его исследования. Корпусная лингвистика обладает как тем, так и другим. В качестве своей главной цели изучаемая нами наука видит объективное лингвистическое описание языковой системы, причём к этому описанию корпусная лингвистика подходит от изучения конкретной человеческой коммуникации, от реальных текстов, которые ранее рассматривались лишь как досадная помеха. В качестве вторичной задачи рассматривается выработка особого способа отражения речевого материала в корпусе текстов. Этот способ, в свою очередь, может использоваться другими лингвистическими дисциплинами. Ещё одно отличие в подходах между традиционной лингвистикой и корпусной заключается в том, что традиционно языкознание изучало возможность (possibility) или невозможность какого-либо лингвистического явления. Например, традиционный учебник английского языка скажет вам, что конструкция I'm not в литературном английском возможна, а конструкция I ain't – нет. Корпусная лингвистика дополнительно изучает и вероятность (probability) лингвистических явлений. То есть, с точки зрения корпусной лингвистики, мы не можем сказать, что употребление I ain't в литературном языке совершенно невозможно. Оно всего лишь маловероятно.
Основные направления корпусной лингвистики.

Кратко и неполно расскажем об основных направлениях современной корпусной лингвистики. Во-первых, это лексикографические исследования, создание словарей. Практически все современные словари английского языка (Collins, Webster, MacMillan и т.д.) издаются на основе огромных корпусов, которые позволяют сделать словарь репрезентативным. То есть, словарь может быть верным или не верным относительно данного корпуса. Во-вторых, изучение корпусов позволяет получать точные данные о лексическом составе языков, об относительных частотах употребления тех или иных слов. В частности, при помощи корпусной лингвистики был окончательно доказан так называемый закон Ципфа, утверждающий, что если в любом естественном языке все слова упорядочить по убыванию частоты их использования, то частота любого слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру (так называемому рангу этого слова). Например второе по частоте слово

встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и так далее.

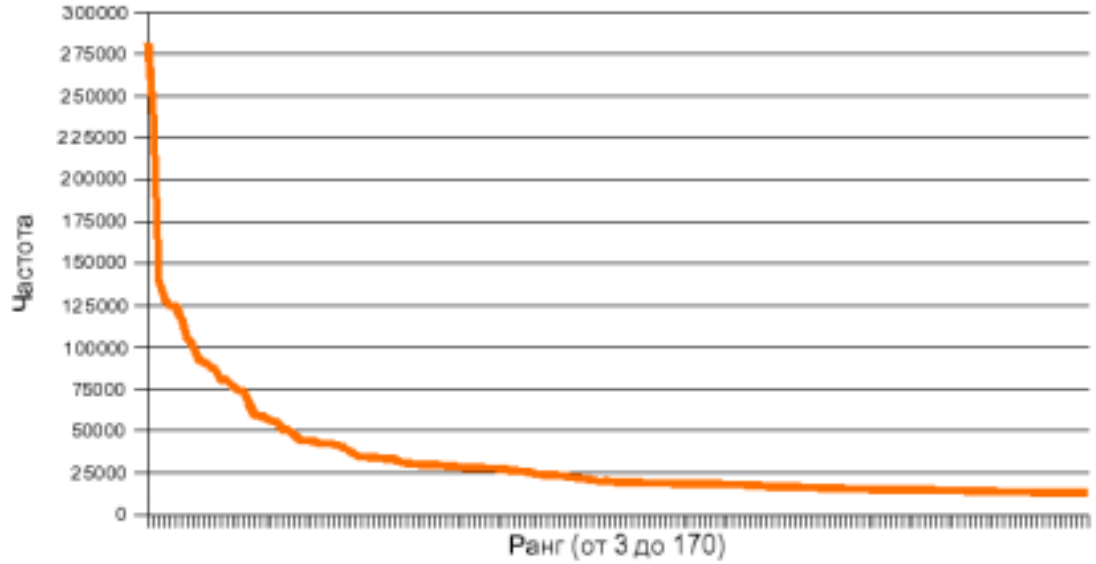

Рисунок 1: Закон Ципфа

Выводом из закона Ципфа является утверждение о том, что язык – это большой набор редких событий. То есть, редких слов в языке значительно больше, чем частых. В-третьих, корпусная лингвистика изучает и изменения в лексическом составе языков, различные его вариации (например, появление и исчезновение неологизмов). Четвёртое направление корпусной лингвистики – изучение грамматики естественных языков, в частности – сочетаемости тех или иных грамматических явлений друг с другом. Естественно, что данные, полученные из живой речи, гораздо более актуальны, чем умозрительные грамматики традиционной лингвистики. Кроме того, получается более объективное исследование: грамматика верна лишь относительно того или иного корпуса текстов. В-пятых, не оставлено без внимания и изучение текстов. Например, используя корпусы, мы можем научиться определять функциональный стиль через статистические характеристики текста – среднюю длину слова и предложения, характерные сочетания слов и т.д. Такие методы уже существуют и используются в автоматическом реферировании и тематическом поиске. Причём, изучать таким образом можно не только письменный, но и устный дискурс.

В-шестых, корпусная лингвистика активно используется в лингводидактике, то есть, в обучении иностранным языкам. Чтобы знать, чему, собственно, учить, необходимы точные количественные данные о преподаваемом языке — состав наиболее частотной лексики, вероятности употребления тех или иных грамматических конструкций и т. д. Что немаловажно, корпусная лингвистика даёт возможность обновить набор примеров, которые используются в преподавании языка. И наконец, особый интерес для нас, как переводоведов, представляют, конечно, многоязычные корпусы, особенно «выровненные» или «сопоставленные» (aligned). В «выровненном корпусе» каждой фразе на одном языке соответствует её эквивалент на другом языке или языках. Такие корпусы используются при подготовке переводчиков или при создании двуязычных словарей. Очень важны они для создания систем автоматического машинного перевода (если такая система опирается на корпус переводов, сделанных переводчиками-людьми, её качество будет гораздо выше). Кроме того, такой корпус можно использовать для исследований, связанных со сравнением оригинальных и переводных текстов.

Корпусная лингвистика и компьютерная лингвистика Довольно часто звучит вопрос о соотношении корпусной и так называемой «компьютерной лингвистики». Эти ветви науки о языке, действительно, близки друг другу, но всё же не совпадают. Что такое «компьютерная лингвистика»? Вообще, термин довольно расплывчат, тем более, что

существует ещё некая «математическая лингвистика». В англоязычном языкознании проще — там есть один общий термин computational linguistics, то есть, «вычислительная лингвистика». Мы для простоты будем говорить «компьютерная лингвистика», поскольку сейчас без компьютеров всё равно никто уже ничего не вычисляет. Так вот, обычно говорят, что компьютерная лингвистика — это такая междисциплинарная ветвь лингвистики, занимающаяся либо статистическим либо rule-based1 моделированием языка с использованием компьютеров. Моделирование – это приблизительный эквивалент английского термина sampling. То есть, компьютерная лингвистика строит модели языка. Кстати, корпусная занимается примерно тем же, поэтому они друг другу помогают.
Вот некоторые точки приложения компьютерной лингвистики:
• автоматический перевод;
• автоматизированное извлечение информации из естественных текстов;
• конструирование удобных интерфейсов между человеком и машиной;
• количественное описание общения на естественных языках;
Немаловажно, что компьютерная лингвистика создаёт инструменты (то есть, программы) для корпусной лингвистики. В этом смысле они тоже дополняют друг друга. Например, корпусным лингвистам необходимы средства для автоматической разметки классов слов в корпусах. Если у вас есть корпус на 100 миллионов словоупотреблений и вам нужно отметить часть речи у каждого слова, то вручную это сделать совершенно нереально. Тут и понадобится специализированное программное обеспечение. Обычно сначала его нужно «обучить», то есть разметить вручную какое-то небольшое количество слов, чтобы система «натренировалась». После этого разметка по классам слов2 будет происходить в автоматическом режиме. Очень активно в современном мире используются программы морфологического и синтаксического анализа 3 . Именно они лежат в основе автоматической проверки орфографии и грамматики, которая в текстовых процессорах подчёркивает вам красным неправильные слова и фразы. Для создания таких программ равно необходимы как программисты, так и лингвисты. Для исследования корпуса бывает важно сначала снять лексическую неоднозначность, то есть, выделить слова-омонимы. Например, в корпусе русских текстов нужно отделить слово «лук» в значении «овощ» от слова «лук» в значении «оружие». В большом корпусе сделать это вручную затруднительно. Поэтому компьютерная лингвистика создаёт программы семантического анализа текстов, которые могут в более или менее автоматическом режиме определять, в каком значении употреблено то или иное слово.
И, наконец, компьютерная лингвистика активно занимается вопросами создания параллельных корпусов, о которых говорилось выше. Ведь это очень интересная лингвистическая задача – как в автоматическом режиме «сопоставить» 1 два текста, один из которых является переводом другого? Как «соотнести» друг с другом отдельные предложения на языке оригинала и на языке перевода? Здесь достаточно проблем и трудностей, но решения уже есть и уже существуют автоматические системы сопоставления текстов. Некоторые из таких программ мы будем изучать в рамках курса «компьютерные технологии в переводе». Итак, как можно видеть, компьютерная лингвистика выступает для корпусной в качестве «поставщика» инструментов анализа и обработки корпусов. Поскольку большой корпус можно обрабатывать только при помощи компьютера, необходимы программы. А написанием лингвистически ориентированных программ как раз и занимается компьютерная лингвистика. С другой стороны, в современной науке порой сложно отделить корпусного лингвистика от компьютерного, поскольку чаще всего учёные занимаются и тем и другим.

Предмет исследования корпусной лингвистики. Развитие лингвистических корпусов в мире: первое и второе поколение.
Предмет исследования Корпусная лингвистика рассматривает текстовые массивы как поле изучения и как источник фактов для лингвистического описания и аргументации. Как уже

говорилось, она сосредотачивается на «речи» (performance), а не на «языке» (competence). Как и вся наука о языке, корпусная лингвистика занимается в основном описанием и объяснением сущности, структуры и использования языка, а так же более частными вопросами: изучение языков, их изменение и т.п. Однако корпусная лингвистика стоит в языкознании несколько особняком Можно отметить, что часто она ограничивается изучением скорее лексики и лексической грамматики, нежели синтаксиса. В чём-то это результат использования методики конкордансов (списков слов в контекстах, в последующих лекциях будет более подробно) – ширины экрана или печатного листа (обычно 130 символов) просто не хватает на то, чтобы анализировать синтаксис или дискурс.

Рисунок 1: Пример конкорданса в программе Corsis

Существует четыре группы корпусных лингвистов: 1. Создатели корпусов (corpus compilers). 2. Разработчики программ для анализа корпусов (corpus software developers) 3. Дескриптивные лингвисты, которые используют существующие корпусы для адекватного описания лексики и грамматики языка. В основном используется вероятностный подход.1 4. Те, кто занимается использованием корпусов в новых прогрессивных приложениях –

2 История электронных лингвистических корпусов
2.1 Первое поколение корпусов
2.1.1 The Brown Corpus
Точное название: Brown University Standard Corpus of Present-Day American English. Составлялся с 1961 по 1964 год. Язык корпуса: американский английский, письменные тексты, 1 миллион словоупотреблений (это количество стало фактическим стандартом для всего первого поколения корпусов). В то время в лингвистике доминировала концепция Хомского, так что Nelson Francis и Henry Kucera (создатели Брауновского корпуса) делали свою работу в очень неблагоприятной атмосфере. Корпус состоит из 500 текстов по 2000 слов каждый. Фактически, он задал стандарт для корпусных исследований, поскольку была очень хорошо продумана структура и выбор категорий текстов. Этот же проект установил традицию свободного доступа к корпусам для исследовательских нужд. На этом корпусе уже в 1969 году был основан словарь American Heritage Dictionary. 2.1.2 Lancaster-Oslo/Bergen (LOB) Corpus 1970-78 год, проект университетов Ланкастера и Осло и научного центра в Бергене. Британский английский, 1 миллион словоупотреблений, структура похожа на Брауновский корпус. Учёные уже начали понимать, однако, что одного миллиона словоупотреблений недостаточно для анализа низкочастотных элементов языка (а их большинство). Тем не менее, на Брауновском и LOB корпусах основаны многие сотни качественных и интересных исследований. Сайт проекта - http://khnt.hit.uib.no/icame/manuals/lobman/ 2.1.3 London-Lund Corpus (LLC) В 1975 году было завершено создание корпуса устной английской речи. Он содержал около 500 тысяч словоупотреблений с орфографической транскрипцией, фонетической и просодической разметкой. Эта грандиозная работа сначала была выполнена в бумажном варианте сотрудниками University College London, а затем переведена в компьютерную форму лингвистами из шведского города Лунд. Сайт проекта - http://www.ucl.ac.uk/english-usage Помимо упомянутых, составлялись корпусы для лексикографических исследований (American Heritage Intermediate), для изучения разговорного английского (Lancaster/IBM

Spoken English Corpus, Corpus of Spoken American English, etc), диахронические корпусы (Helsinki Corpus of English Texts: Diachronic Part, 1,5 миллиона словоупотреблений), корпусы для лингводидактических исследований (International Corpus of Learner's English) и другие.

2.1.4 Машинный Фонд русского языка Создание первого советского лингвистического корпуса началось в 1985 году в Институте русского языка Академии Наук СССР. Успели только разработать концепцию и архитектуру корпуса и несколько программ, а также собрать какое-то количество текстов. В районе 1991 года финансирование прекратилось и работы заглохли.

2.1.5 Уппсальский корпус русского языка Одновременно (в 1980-е годы) в институте славистики университета Уппсалы (Швеция) был создан Уппсальский корпус современных русских текстов. 1 миллион словоупотреблений, около 600 текстов. Сайт проекта - http://www.slaviska.uu.se/korpus.htm

2.2 Второе поколение корпусов К 90-м годам технологии извлечения и хранения текстов позволили создавать корпусы из ста миллионов словоупотреблений и более.

2.2.1 The Cobuild Project / The Bank of English Началось всё в 1980 году, когда издательство Collins принялось за составление корпуса для создания нового словаря. В 1990 году было объявлено об объединении усилий Collins и факультета английского языка университета Бирмингема в инициативу под названием The Bank of English. The Bank of English — это так называемый мониторинговый корпус. Вот слова руководителя проекта Джона Синклера: мониторинговый корпус это огромный, вечно изменяющийся поток языка, не имеющий чётко определённого размера. Этот поток проходит через фильтры, которые извлекают из него лингвистические данные. Около 300 миллионов словоупотреблений в 1997 году, а в 2005 уже 525 миллионов. Каждый месяц в корпус поступает два с половиной миллиона новых словоупотреблений. 25 процентов корпуса составляет устная речь, 75 процентов — письменная.

По адресу http://www.collins.co.uk/Corpus/CorpusSearch.aspx можно использовать тестовую версию корпуса (56 миллионов словоупотреблений).

2.2.2 The Longman Corpus Network Коммерческая база данных нескольких корпусов, созданных компанией Longman и университетом Ланкастера. 50-100 миллионов словоупотреблений. Сайт в Интернете - http://www.pearsonlongman.com/dictionaries/corpus/index.html

2.2.3 British National Corpus (BNC) 100 миллионов словоупотреблений, представляет английский язык в целом, а не один жанр. Этот корпус имеет конечный размер, в отличие от Cobuild Project. 90 процентов письменных текстов, 10 устных. В создании принимали участие многие организации, включая Британское правительство. Процесс завершился около 1995 года. Корпус состоит из 4124 текстов, из которых 863 транскрибированы из устных бесед или монологов. Каждый текст сегментирован на орфографические предложения, а внутри них каждому слову автоматически назначен код класса слова (части речи). Во всём корпусе 6,4 миллионов орфографических предложений. Сегментирование и классификация слов были выполнены программой стохастической разметки CLAWS, разработанной в университете Ланкастера. Классификационная схема предусматривает 65 частей речи, которые описаны в прилагающейся документации. Все тексты размечены с использованием наиболее стандартных способов — языка SGML и системы TEI. При создании корпуса были использованы новые подходы к отбору текстов, многоуровневая система контроля. Корпус доступен по адресу http://www.natcorp.ox.ac.uk/

2.2.4 The International Corpus of English (ICE) Совместный проект нескольких десятков университетов. 20 параллельных подкорпусов, по миллиону словоупотреблений каждый, вместе 20 миллионов словоупотреблений. Можно изучать специфику стран, где английский – второй или официальный язык (Австралия, Канада, Новая Зеландия и т.п.). Разработано сложное программное обеспечение специально для анализа этого корпуса. Веб-сайт: http://www.ucl.ac.uk/english-usage/ice/

2.2.5 American National Corpus Первый выпуск состоялся в 2003 году. Планируется 100 миллионов словоупотреблений, но пока только 11. Доступ исключительно на платной основе. Корпус в XML-формате. Веб-сайт проекта: http://www.americannationalcorpus.org/
2.2.6 Gigaword corpora Мониторинговые корпусы английского, арабского, китайского и других языков. Спонсируются Европейским Союзом, создаёт их компания Linguistic Data Consortium. Уже 1 миллиард словоупотреблений. В основном тексты взяты из публицистики и новостей. Корпусы довольно дорогие. Посмотреть на список можно на сайте http://www.ldc.upenn.edu
2.3 Современные российские лингвистические корпусы
2.3.1 Национальный корпус русского языка Общедоступный для поиска корпус русских текстов (сокращённо НКРЯ). Открыт 29 апреля 2004 в Интернете по адресу http://ruscorpora.ru. Работы по созданию Корпуса были начаты в 2001 году группой лингвистов из Москвы, Петербурга, Воронежа и других городов. Основные участники – Институт русского языка РАН, Институт языкознания РАН и компания «Яндекс». Письменные, устные, поэтические диалектные тексты. 140 миллионов словоупотреблений в 2007 году. Корпус морфологически и семантически размечен и полностью свободен для использования при помощи веб-сайта.

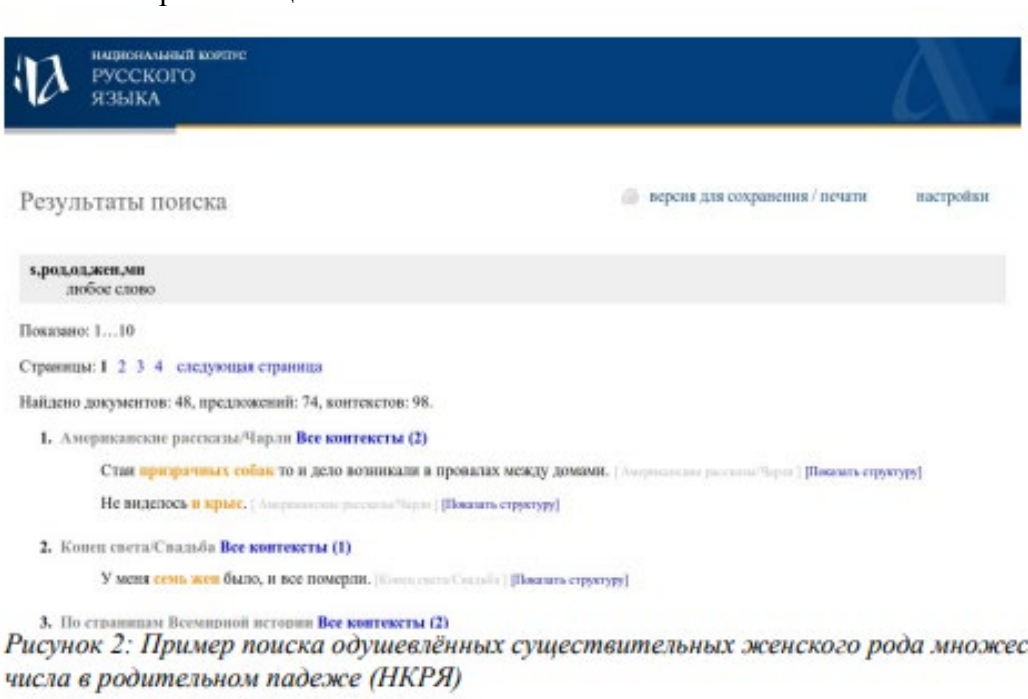

Рисунок 2: Пример поиска одушевлённых существительных женского рода множественного числа в родительном падеже (НКРЯ)

Корпусы: устные и письменные, одноязычные и многоязычные
Если бы нам пришла в голову идея исследовать корпус текстов по корпусной лингвистике (например, книг и научных статей) методами самой корпусной лингвистики, то оказалось бы, что чаще всего к слову «корпус» примыкает глагол «составлять»
1 . Какими же бывают корпусы по методу составления? Устные – письменные Большая часть корпусов 1 поколения были исключительно письменными. Письменные тексты гораздо легче собирать.
Существуют три метода ввода письменных текстов в компьютер:
● заново набирать тексты (это лучше, чем пробивать перфокарты, как было с Брауновским корпусом);
● использовать тексты, которые уже существуют в электронной форме;
● сканировать напечатанные тексты (но при этом нужно исправлять много ошибок). Большие современные корпусы обычно комбинированные, с преобладанием письменных текстов. Даже в BNC лишь 10% текстов устные. Выделяется ICE, в котором 60% текстов устные. Между тем, язык в основном существует именно в устной форме, письменная его

форма вторична. Поэтому так важны устные корпусы, либо смешанные. Среди специфически устных корпусов нужно назвать London Lund Corpus (LLC, 1975 г.) и Lancaster/IBM Spoken English Corpus (1992), сокращённо SEC. Этот последний состоит из 52600 словоупотреблений. Он поставляется на CD-ROMe вместе с аудиозаписями, полностью размечен на предмет ударений, интонации, пауз и т.п. Однако, он не содержит информации о социальном статусе и образовании респондентов, что ограничивает его использование в социолингвистике. Corpus of Spoken American English (1991), миллион словоупотреблений, 80 часов звучания. Map Task Corpus (1991, университет Глазго, Шотландия), 147 тысяч словоупотреблений, 16 часов звучания.

Устные корпусы включают меньше словоупотреблений, чем письменные, не только изза трудоёмкости сбора данных, но и потому, что для просодических исследований обычно достаточно меньшего количества слов. Так, для изучения интонации достаточно корпуса в сто тысяч словоупотреблений. Устные корпусы могут включать как монологическую, так и диалогическую речь. Для сбора материала используются записи с радио и телевидения или опрос по выборочным методикам социологии и социолингвистики. Отметим, что скрытая запись сейчас считается неэтичной (в отличие от 70-х годов). Обычно собирают довольно подробную информацию о респондентах: ● место записи ● что респондент делает ● время ● дата ● количество участников ● степень спонтанности беседы ● тема ● пол участников ● возраст участников ● этническая принадлежность участников ● основной язык участников ● профессия ● образование ● социальный статус ● отношение к записывающему ● диалект.

Самая трудоёмкая стадия — transcription. Орфографическая транскрипция одного часа записи с минимальной интонационной разметкой может занять около 10 часов. Если же размечать текст по всем правилам TEI (Text Encoding Initiative), то на это может уйти 25 часов и более. А без разметки корпус устных текстов не имеет смысла — как минимум, должна быть указана продолжительность пауз, размечена одновременная речь, ударение, интонация. Иногда включают контекстные комментарии типа «ест печенье». Именно благодаря подробной разметке корпус LLC стал стандартном для корпусов устной речи.

Статические – динамические

Первые корпусы были статичными снимками языка. Наиболее значимый современный корпус (BNC) тоже статичен. Но начали появляться и динамические мониторинговые корпуса, которые пополняются постоянно. Пример — Cobuild Project. Такие корпусы ещё называются «открытые». Их проблема в том, что они часто не совсем адекватно представляют язык, поскольку не подчиняются чётким критериям отбора, тексты не сбалансированы.

Одноязычные — многоязычные

Корпусных лингвистов (особенно связанных с переводом) всегда интересовала задача составления корпусов на нескольких языках. Уже в первом поколении начали появляться двуязычные корпусы для таких языков, как английский, финский, французский, немецкий, греческий, норвежский, испанский, шведский, валлийский. Такие корпусы ещё называются bitexts.

Естественно, нет никаких технических препятствий к тому, чтобы делать корпусы не дву- а трёх-, четырёх- и более язычными. Вообще говоря, само появление многоязычных корпусов спровоцировало всплеск научных исследований, поскольку для их анализа требуются другие инструменты и даже другие концепции, нежели чем для анализа корпусов одноязычных.

Вполне естественно, что можно представить себе два типа двуязычных корпусов: • корпус, в котором тексты являются переводами друг друга • корпус, в котором просто присутствуют тексты на разных языках (возможно, одной и той же тематики). Корпусы второго типа иногда называют «переводными» (translation corpora) и используются для изучения различий в выражении схожих мыслей на разных языках. Корпусы первого типа называют «параллельными» (parallel corpora) и используются для исследования различных

аспектов собственно перевода. Например, существует параллельный корпус текстов заседаний канадского парламента (английский/французский). Параллельные корпусы также могут быть двух типов — выровненные (aligned) и не выровненные (not aligned). «Выровненность» означает, что в корпусе существует чёткая связь между единицами перевода, которые соответствуют друг другу. То есть, мы можем быстро найти, как то или иное слово или предложение переводилось на другой язык. Обычно такими единицами перевода служат всё-таки предложения, поскольку часто сложно выровнять слова (ведь обычно переводят не дословно). Такой корпус наиболее полезен для переводчика, поскольку представляет собой ту самую «память переводов» (translation memory) — бесценный ресурс, позволяющий использовать предыдущие переводы. Невыровненные корпусы ещё называют «сравнительными». «Выровнять текст с его переводом на другой язык означает показать какие части текста переведены какими частями второго текста» (Kay & Röscheisen 1993: 121) Выравнивание (alignment) можно делать автоматически, а можно вручную. Первый способ быстрее, но чреват ошибками. Например, если при переводе произошло членение или объединение предложений, то не всегда можно легко определить, какое из предложений перевода соответствует какому предложению оригинала. Одним из примеров выровненного многоязычного корпуса может послужить база данных Acquis Communautaire Европейского Союза (DGT-TM). Это память переводов европейского законодательства на 22 языка1 , которую выложили в открытый доступ в ноябре 2007 года. Всего в ней около миллиарда слов, она выровнена по предложениям (sentencealigned). Вот пример предложения из этой базы данных:
EN: Articles 5 to 7 of this Directive do not apply to containers for gases which are compressed, liquefied or dissolved under pressure.
BG: Членове 5 - 7 на настоящата директива не се отнасят за контейнери с газове, които са сгъстени, втечнени или разтворени под налягане.
CS: Články 5 až 7 této směrnice se nevztahují na kontejnery pro plyny, které jsou stlačené, zkapalněné nebo rozpuštěné pod tlakem. и т.д.
Ценность параллельного корпуса, как и других корпусов, возрастает с его размером и количеством языков. В этой связи трудно переоценить важность Acquis Communautaire, который является самым большим параллельным корпусом в мире. Ещё два его преимущества — бесплатность и наличие редких пар языков, типа «мальтийский-эстонский», «словенский финский». Сам корпус представлен в стандартном открытом формате памяти переводов TMX, про который я ещё расскажу на лекции по компьютерным технологиям в переводе. Этот и подобные корпусы можно использовать для многих целей. Например: ● выявление типичных переводческих приёмов и трансформаций ● обучение статистических систем автоматического перевода ● создание одноязычных и многоязычных словарей ● обучение и тестирование программ извлечения информации ● автоматическая проверка правильности перевода ● и конечно, облегчение труда переводчика через подбор возможных эквивалентов Двуязычные корпусы — ещё одно благодатное поле для студентов-лингвистов, которые могут использовать их для выполнения своих квалификационных работ. Корпус в данном случае может пониматься не как самостоятельная цель, а как инструмент для получения некоторых языковых данных. Соответственно, здесь возможны либо исследования процесса и результата перевода (берём оригинал и перевод), либо контрастивные исследования (берём схожие тексты на языке 1 и языке 2).

Корпусы: аннотированные и неаннотированные. Лингвистическая аннотация (разметка) и метаданные.
Что такое разметка? Знаки пунктуации — это разметка. Маргиналии на полях средневековых манускриптов — это разметка. Под лингвистической аннотацией или разметкой корпуса (по-английски linguistic markup) подразумевается наличие в корпусе неких данных, не являющихся частью текста, но несущих какую-то информацию о нём

(так называемые метаданные). Простейший пример таких данных — отметки частей речи. Выглядеть это может так: I will use Google before asking dumb questions. Размечаем: I (pronoun)will (verb) use (verb) Google (noun) before (preposition) asking (verb) dumb (adjective) questions (noun) . В основном это нужно для облегчения автоматического анализа корпуса. Один раз отметив в тексте все части речи, затем можно производить любые исследования, связанные с ними без необходимости заново выявлять, например, все прилагательные в корпусе. Понятно, что если такой разметки нет, то, к примеру, поиск по слову «will» выдавал бы все случаи его появления в корпусе, вне зависимости от того, существительное это или вспомогательный глагол. Но ведь обычно исследователя интересует лишь какой-то один из этих случаев! И это далеко не единственный тип разметки, который бывает нужен корпусному лингвисту.

История систем разметки

В 80-х годах был принят стандарт разметки электронных текстов под названием SGML1 (Standard Generalized Markup Language).Он был разработан внутри типографской индустрии, но быстро распространился на другие отрасли. Смысл SGML был в том, чтобы документы, набранные в разных текстовых процессорах, можно было редактировать, анализировать и изменять в любом из них.

Тэги как лингвистический инструмент

SGML ввёл концепцию тэгов. Тэги (англ. tags) — это служебные пометки в тексте, содержащие информацию о самом тексте. Для каждого случая можно определять собственные тэги и таким образом создавать диалекты языка SGML. Традиционно тэги заключаются в угловые скобки и бывают парными: открывающими и закрывающими. Например, - это открывающий тэг, а - закрывающий. Закрывающий тэг сигнализирует, что то, о чём сообщал открывающий тэг, закончилось. Приведём пример тэга *(выделение важного в тексте, emphasis): Это относится в первую очередь к вам! В данном случае слова «в первую очередь» помечены как важные. Тэги могут быть вложенными друг в друга: Это относится в первую очередь к вам! - сказал он. Текст «это относится в первую очередь к вам» заключён в тэги , означающие прямую речь (direct speech), а внутри него слова «в первую очередь» дополнительно заключены в тэги . Количество уровней вложенности не ограничено. Тэги могут быть и не парными, то есть, не иметь «открывающей» и «закрывающей» части. Например, при разметке устных корпусов употребляется тэг , означающий, что в этом месте произошла задержка речи. Он одиночный. Сами тэги в обычных обстоятельствах пользователю не показываются. Программа, отображающая размеченный текст, интерпретирует тэги в соответствии с заложенными в неё правилами и показывает пользователю текст, оформленный согласно им.*

Текст с разных сторон: alternative views

Одно из наиболее значительных преимуществ разметок семейства SGML — возможность нескольких представлений текста (alternative views). Это означает, что один и тот же размеченный текст легко представить в нескольких видах, в зависимости от нашей текущей задачи. Например, мы хотим выделить из корпуса только текст, не являющийся прямой речью. Тогда та программа, в которой мы просматриваем текст, просто скроет все символы, заключённые в тэги и наш пример будет выглядеть уже так: - сказал он. Или мы можем указать, чтобы текст, помеченный, как важный, был зелёного цвета, а прямая речь выделялась полужирным шрифтом: Это относится в первую очередь к вам! - сказал он.

Потомки SGML

Язык разметки SGML — это как бы «конструктор» языков. Сам по себе, в своём первозданном виде, он очень сложен и используется довольно редко. Но на его базе были созданы такие широко известные языки разметки, как HTML и XML. Язык HTML (Hyper-Text Markup Language), на котором написано подавляющее большинство страничек интернет-сайтов, создали из SGML путём выделения чётко определённого ограниченного набора тэгов, в основном относящихся к оформлению, а не к содержанию документа. В

результате мы получили WWW (Всемирную Паутину). Второе широко известное подмножество SGML — расширяемый язык разметки XML (eXtensible Markup Language), который применяется для хранения любых структурированных данных — в том числе и текстов в корпусах. Фактически, это свод синтаксических правил для описания структуры данных. Например, формат офисных документов Open Document построен именно на XML. Специально для разметки текстовых данных (корпусов) несколько университетов2 разработали систему, описывающую, какие именно параметры текстов нужно размечать. Эта система использует XML и называется Text Encoding Initiative Guidelines (TEI Guidelines). Это список различных особенностей текстов, которые вообще можно кодировать, размечать и индексировать. Например, система перечисляет различные типы исправлений в тексте, помарок, цитат, иностранных слов и т.д. и т.п. В настоящее время практически все проекты по созданию корпусов (в том числе British National Corpus) стараются в той или иной мере следовать рекомендациям TEI. Подробнее почитать о них можно на http://www.teic.org/Guidelines/index.xml. Естественно, каждый, кто создаёт корпус, может сам выбирать, что именно ему размечать и насколько подробно. Но считается, что в письменном корпусе нужно размечать части речи, границы высказываний, цитаты, списки, заголовки, аббревиатуры, имена собственные, инициалы и акронимы, главы книг. В устных текстах важно разметить обмен репликами, прерывания, перекрывающуюся речь, диалектные формы, паузы и неразличимую речь. В приложении к этой лекции приведён пример текста, размеченного в соответствии с рекомендациями TEI.

Автоматическая разметка текстов

Понятно, что размечать большие корпусы вручную — занятие очень долгое и дорогое. Поэтому уже в 70-х годах появляются первые проекты по поручению этой задачи компьютеру. Тогда программа TAGGIT смогла корректно назначить тэги частей речи 77% слов в Брауновском корпусе. Остальные пришлось размечать вручную в течение 10 лет. Но прогресс не стоял на месте. В 80-е годы система CLAWS (Constituent Likelihood Automatic Word-tagging System) правильно разметила уже около 95% Брауновского корпуса. В ней использовался вероятностный подход. В настоящее время для основных европейских языков уже реализованы как автоматическая разметка частей речи (морфологический анализ, word-class tagging), так и автоматическая разметка членов предложения (синтаксический анализ, parsing). Эти достижения используются, в том числе, и в системах автоматического перевода и интернетпоиска. В этой связи нужно отметить немалый вклад рабочей группы учёных под названием «Автоматическая обработка текста» (сайт http://www.aot.ru). В основном они занимаются русским языком. Выросла эта группа из факультета лингвистики РГГУ и занимается приложением теоретической лингвистики к современным компьютерным технологиям. Они разработали модули графематического (определение границ слов), морфологического (определение частей речи), синтаксического (определение членов предложения) и семантического (выявление семантических связей между словами) анализа текстов на русском, немецком и английском языках.

Лингвистические исследования на базе корпусов

Лингвистические корпусы составляют, чтобы предоставить основу для более точного и адекватного описания структурных и функциональных параметров языка. Сегодня мы поговорим о результатах некоторых корпусных исследований и опишем как вообще использование корпусов может помочь лингвистике.

Описания лексики

Конечно, чаще всего корпусные описания лексики применяются в лексикографии. Практически все современные словари английского языка построены на базе корпусов. Корпусы помогают достоверно определить набор словоформ (types) в языке, показывают появление новых словоформ, используются для уточнения разных значений одного слова

и их относительных частот. Самый известный пример — словари издательства Collins, построенные на базе корпуса Cobuild Project. Корпусы могут давать очень интересные лексикографические сведения о языке. Так, даже в относительно небольшом London-Lund Corpus слово good встречается 800 раз. Оно выступает в 20 значениях как прилагательное, а кроме того, может являться междометием в различных функциях. У всех этих значений разная частота употребления. Кроме того, корпусы показывают появление неологизмов. Так, в 1994 году в английских газетах появились следующие интересные слова: complainy, dial-a-video, bespokesman, cleavage-wielding, eventdriven, fruitcakeland, infotainers, over-housed, unbusy, anarchitecture, bimboisation, bonkable, crashworthiness. Статистические исследования лексики на материале корпусов начались ещё в докомпьютерную эпоху. Основным их результатом стал известный закон Ципфа (30-е годы). Напомним, что суть его в том, что в любом массиве текстов небольшое число словоформ1 (types) образует большую часть реальных словоупотреблений2 (tokens). Соответственно, например, 90-95 процентов словоупотреблений в английских текстах составлено из 2-5 тысяч наиболее употребительных словоформ. Более того, около половины текста — это словоупотребления 50-100 самых актуальных словоформ (хотя конкретный их набор может быть разным для разных стилей и подъязыков). Это открытие имело большое значение для преподавания английского языка, поскольку позволило сосредоточиться на предъявлении учащимся самой частотной лексики. Существуют исследования, которые описывают, какая лексика специфична для определённых типов текстов и вряд ли появится в других. Так, для научных текстов одним из таких слов является глагол to measure, а для художественных — to kiss. Появление электронных корпусов дало возможность уточнить частотные параметры лексики. Вот, например, список 50 наиболее частых словоформ в Birmingham Corpus:

1) the	*14) you*	*27) are*	*40) so*
2) of	*15) on*	*28) or*	*41) what*
3) and	*16) with*	*29) by*	*42) their*
4) to	*17) as*	*30) we*	*43) if*
5) a	*18) be*	*31) she*	*44) would*
6) in	*19) had*	*32) from*	*45) about*
7) that	*20) but*	*33) one*	*46) no*
8) I	*21) they*	*34) all*	*47) said*
9) it	*22) at*	*35) there*	*48) up*
10) was	*23) his*	*36) her*	*49) when*
11) is	*24) have*	*37) were*	*50) been*
12) he	*25) not*	*38) which*	
13) for	*26) this*	*39) an*	

http://lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf

2. МЕТОДИЧЕСКИЕ УКАЗАНИЯ

2.1. МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ ЗАНЯТИЯМ

Целью практических занятий является развитие у студентов навыков лингвистического анализа, что будет способствовать также и усвоению теоретических знаний. Выполнение заданий на практических занятиях может принести пользу только в том случае, если решение лингвистических задач потребует от студента известных усилий (обращение к рекомендуемым библиографическим источникам, словарям, справочникам, ресурсам интернет).

2.1.1. ДОКЛАД ПО ТЕМЕ ЛЕКЦИИ

В ходе лекций студенты получают новые теоретические знания по изучаемой дисциплине. Вопросы, не рассмотренные/рассмотренные в сжатом виде на лекции, выносятся на лабораторные занятия и предварительно должны быть изучены студентами самостоятельно. На лабораторном занятии каждый его участник должен быть готов к выступлению по всем поставленным в плане вопросам, проявлять максимальную активность при их рассмотрении. Выступление должно строиться свободно, убедительно и аргументировано. Преподаватель следит, чтобы выступление не сводилось к репродуктивному уровню (простому воспроизведению текста), не допускается и простое чтение конспекта. Необходимо, чтобы выступающий демонстрировал собственное отношение к тому, о чем он говорит, высказывал свое личное мнение, понимание, обосновывал его и мог сделать правильные выводы из сказанного. При этом студент может обращаться к записям конспекта и лекций, непосредственно к первоисточникам, использовать знание художественной литературы и искусства, факты и наблюдения современной жизни и т. д.

Доклад по теме лекции представляет собой устный ответ, подготовленный на основе конспекта прослушанной лекции и/или конспекта предложенной преподавателем основной и дополнительной литературы. Доклад может сопровождаться электронной презентацией. Продолжительность доклада составляет 7-10 минут, по его окончании студенты и преподаватель задают докладчику вопросы по теме выступления (5 мин.). Доклад должен содержать следующие компоненты:

- тема доклада;
- план доклада;
- основные определения в представляемой области;
- основной текст доклада;
- выводы по теме.

Доклад должен быть изложен научным стилем. Не допускается использование: длинных сложных предложений, затрудняющих восприятие; малоупотребительных иностранных слов, узкоспециальной терминологии, известной ограниченному кругу профессионалов (без объяснения их значений); вводных конструкций, не несущих смысловой нагрузки; общих слов. Позиция автора в докладе должна демонстрироваться минимально, недопустимо использование местоимений «я», «моя» (точка зрения).

Темы докладов по разделу. **Лексикография как наука. Словарь как способ семантизации общей и специальной лексики в прикладных целях.**

Задание. Разработайте карты-конспекты следующих вопросов. На основе Ваших карт подготовьте презентацию. Перескажите на занятии изученный материал устно.

1. Лексикография - древнейшая область лингвистического описания.
2. Этапы развития практической и теоретической лексикографии.
3. Предмет, цели и задачи современной теоретической лексикографии.
4. Определение понятия «словарь». Словарь как способ семантизации общей и специальной лексики в прикладных целях.
5. Принципы и функции универсального словаря.
6. Макроструктура словаря. Микроструктура словаря.

7. Вводная часть словаря: концепция словаря и ее обоснование.
8. Метаязык словаря. Лексикографический параметр.

Темы докладов по разделу. **Типология словарей**. **Типологические параметры словаря. Современные классификации словарей.**
Задание. Разработайте карты-конспекты следующих вопросов. На основе Ваших карт подготовьте презентацию. Перескажите на занятии изученный материал устно.
1. Адресат, универсальные словарные функции, источники, вход в словарь, входная словарная единица, объем словаря, формат словаря, макроструктура, микроструктура, средства семантизации.
2. Понятия *тип, типология, классификация*. Основополагающая типология словарей Л.В.Щербы.
3. Классификации по пересекающимся и непересекающимся признакам.
4. Фасетные классификации.
5. Тип словаря как набор лексикографических параметров.
6. Параметрическое описание словаря.
7. Определение типа словаря. Типы словарей. Толковые словари.
8. Функции, структура, типологические параметры словарей.
9. Способы и средства семантизации: определение, толкование, описание, экзэмплификация, отсылка, иллюстрация.

Темы докладов по разделу. **Типы словарей. Идеографические словари. Специальные словари.**
Задание. Разработайте карты-конспекты следующих вопросов. На основе Ваших карт подготовьте презентацию. Перескажите на занятии изученный материал устно.
1. Практическая и теоретическая идеография.
2. Словари-тезаурусы**.**
3. Типы идеографических словарей. Словари идеографического типа.
4. Типологические параметры идеографических словарей. Идеографический характер современных словарей.
5. Терминоведение как основа научно-технической лексикографии.
6. Термин – объект лексикографического описания.
7. Проблематика определения термина в словарях разных типов.
8. Способы и средства семантизации термина в специальном словаре.
9. Роль лингвиста в лексикографическом описании термина.

Темы докладов по разделу. **Типы терминологических словарей. Основные понятия электронной лексикографии.**
Задание. Разработайте карты-конспекты следующих вопросов. На основе Ваших карт подготовьте презентацию. Перескажите на занятии изученный материал устно.
1. Функциональные типы специальных словарей: словари-справочники, систематизирующие словари, специальные толковые словари, переводные терминологические словари, тезаурусы, глоссарии.
2. Терминологические словари в электронном формате.
3. Принципы классификации терминологических словарей.
4. Моделирование терминологических словарей-глоссариев**.**
5. Типологические параметры словарей-глоссариев.
6. Принципы и этапы моделирования систематизирующих глоссариев.
7. Типологические параметры и функции переводных терминологических словарей.
8. Принципы моделирования переводных терминологических словарей.
9. Лексикографические параметры электронных словарей разных типов.
10. Макро- и микроструктурные параметры электронных словарей.

11. Учебные англоязычные электронные словари.
12. Переводные электронные словари и онлайновые ресурсы переводчика.

Темы докладов по разделу. **Корпусная лингвистика: предмет, терминология. Связь корпусной лингвистики с др. науками. Методы корпусной лингвистики.**
Задание. Разработайте карты-конспекты следующих вопросов. На основе Ваших карт подготовьте презентацию. Перескажите на занятии изученный материал устно.
1. Корпус как поисковая система.
2. Лингвистические (языковые) и нелингвистические корпусы.
3. История лингвистических корпусов: от картотеки к корпусу.
4. Корпусная лингвистика: современное состояние.
5. Корпусная лингвистика в России.
6. Корпусоподобные интерфейсы между лингвистом и поисковыми системами Интернета.
7. Лингвистические исследования, базирующиеся на корпусах.
8. Создание электронной хрестоматии по корпусной лингвистике.
9. Исследование механизмов взаимодействия корпуса текстов и электронной картотеки (корпусы цитат).
10. Создание веб-сайта по корпусной лингвистике.
11. Предварительные работы по созданию корпуса.
12. Проблемы репрезентативности. Отбор источников. Внешние и внутренние критерии отбора. Нормализация файлов. Графематический анализ.
13. Средства создания и разметки корпусов. Понятие разметки. Типы разметки. Автоматический морфологический и синтаксический анализ. Металингвистическая разметка. Параллельные корпусы. Проблема выравнивания.
14. Стандартизация в корпусной лингвистике. Языковые средства представления размеченных текстов. Международные стандарты и проекты (TEI, EAGLES, CDIF, XCES).

Темы докладов по разделу. **Способы использования корпусов. Типы корпусов.**
Задание. Разработайте карты-конспекты следующих вопросов. На основе Ваших карт подготовьте презентацию. Перескажите на занятии изученный материал устно.
1. Способы использования корпусов в лингвистических исследованиях.
2. Исследование способов использования корпусов в лексикографии.
3. Изучение средств обработки корпусных данных, представленных на языке XML.
4. Обзор существующих корпусов различных типов.
5. Классификация (типология) корпусов по различным основаниям.
6. Типы корпусов по задачам. Типы корпусов по формальным признакам.

2.1.2. ТЕРМИНОЛОГИЧЕСКИЙ ДИКТАНТ

Терминологический диктант – процедура, при которой испытуемому необходимо определить термин по его толкованию или, наоборот, дать определение предложенному термину. Терминологический диктант, как правило, занимает 10-15 минут занятия и проводится в его начале или конце. Термины и определения, содержащиеся в диктанте, соответствуют пройденной теме, содержатся в конспектах лекций по дисциплине и в глоссариях, которые составляются студентами индивидуально.

Образец терминологического диктанта
1. Лексикография – это …..
2. Лингвистический корпус – это ….
3. Корпусная разметка – это …
4. Графематический анализ – это …
5. Писковая система – это …

2.1.3. ТЕСТ

Тестовые задания предназначены для проведения текущего и итогового контроля усвоения содержания дисциплины. Используются следующие формы тестовых заданий: открытая, закрытая (с выбором одного или нескольких правильных ответов), на установление соответствия и последовательности, на дополнение.

При выполнении тестов студенту, прежде всего, рекомендуется внимательно прочитать задание, ответить на вопрос, что необходимо сделать. Чтобы правильно выполнить задание закрытой формы (отметить один или более правильных ответов), необходимо прочитать тестовое утверждение и в приведенном списке отметить сначала те ответы, в которых студент уверен, и определить те, которые точно являются ошибочными, затем еще раз прочитать оставшиеся варианты, подумать, не являются ли еще какие-то из них правильными. Важно дочитать варианты ответов до конца, чтобы различить близкие по форме, но разные по содержанию ответы.

Образцы тестов по дисциплине «Корпусная лингвистика»

Тестовые задания по теме: «Лексикография как наука. Словарь как способ семантизации общей и специальной лексики в прикладных целях»

1. Что такое лексикография?
а) искусство составления словарей;
б) *наука о словарях и практике их составления;*
в) наука о формулировании понятий в словарях;
г) наука о словарной терминосистеме.
2. В энциклопедических словарях описываются и разъясняются …
а) слова, которые называют понятия;
б) *явления, которые словами названы;*
в) междометия, местоимения, служебные слова;
г) наречия, прилагательные, глаголы, которые не являются специальными терминами.
3. Языковые словари показывают …
а) *слова с их значениями, употреблением, происхождением, грамматической характеристикой и фонетическим обликом;*
б) произнесение, обусловленное нормой;
в) термины с их определениями;
г) падежные окончания слов
4. .Одноязычные словари – это словари…
а) *толковые, в задачу которых входит не перевод, а характеристика данного слова в современном языке;*
б) переводные, где дается перевод на родственные титульному языку;
в) терминологические с формулировкой профессиональных понятий;
г) справочные по любым областям знаний.
5. Словари литературного языка – это словари, где…
а) даются примеры из художественной литературы разных эпох;
б) *диалектизмы и областные слова встречаются лишь только в тех случаях, когда они отмечены в литературных памятниках;*
в) показываются термины и метаязык литературы изучаемого языка;
г) не показывается правильное и неправильное употребление слов, их грамматические изменения и произношение.
6. Существуют словари иностранных слов, где даются …
а) перевод иностранного слова с языка-источника;
б) критические замечания о неверном использовании иностранных слов;
в) объяснения, почему заимствовано то или иное слово;
г) *толкования только заимствованных слов.*
7. В двуязычных переводных словарях …
а) *наряду с краткими лексикологическими и грамматическими указаниями к вокабуле*

(заглавному слову) дается перевод данного слова в разных его значениях на другой язык;
б) имеются критические замечания о неверном использовании иностранных слов;
в) имеются термины с формулировкой профессиональных понятий;
г) наряду с краткими лексикологическими и грамматическими указаниями, даются примеры употребления слова.
8. Имеют место быть «обратные словари», где …
а) слова напечатаны, начиная с последней буквы;
б) *слова расположены не в порядке начальных букв, а в порядке конечных;*
в) представлен грамматический состав слов, начиная с последней морфемы;
г) показана синтаксическая функция каждого слова.
9. Автором «Обратного словаря современного русского языка» является …
а) М.И. Откупщикова;
б) О.Н. Гринбаум;
в) Х. Х. Бильфельдт;
г) А.С. Герд.
10. Из чего не состоит словарь:
а) словника, т. е. подбора вокабул (заглавных слов, в немецкой лексикологии это называется Stichwörter) со взаимными ссылками и отсылками,
б) филиации, т. е. расчлененной подачи значений той или иной вокабулы,
в) стилистических, грамматических и фонетических ремарок или помет к словам и их значениям,
г) иллюстративных примеров,
д) идиоматических и фразеологических сочетаний к данному слову;
е) перевода (в разноязычных словарях) или толкования (объяснения – в одноязычных словарях);
ж) *синтаксических функций употребления слова.*
11. К активным формам информационных ресурсов не относят …
а) алгоритмы, б) *базы данных*, в) программы, г) базы знаний
12. К словарям, отражающим некоторые тематические и стилевые пласты лексики не относятся …
а) терминологические словари, б) словари арго, в) *словари архаизмов,* г) диалектные словари
13. Паронимические словари это словари слов, имеющих …
а) *с частичным звуковым сходством при семантическом различии*; б) с полным звуковым сходством при семантическом различии; в) с частичным звуковым различием при частичном семантическом сходстве; г) с полным звуковым различием при частичном семантическом сходстве
14. Каждая словоформа текста характеризуется численными показателями (частотой, номером страницы, номером строки и т.д.) и контекстом в …
а) идеографических словарях; б) *конкордансах;* в) словоуказателях; г) этимологических словарях
15. Теоретической основой контекстологических словарей является теория …
а) звукоподражаний; б) вероятности; в) *детерминант*; г) фразеологических единиц
16. … не является номеном.
а) *кафе*; б) Мак Дональдс; в) «Apple»; г) Атлантический океан
17. Энциклопедии отличаются от словарей тем, что …
а) в первых больше слов; б) во вторых больше слов; в) *первые содержат характеристики не слова как такового, а обозначенного им предмета, факта или явления;* г) вторые содержат характеристики не слова как такового, а обозначенного им предмета, факта или явления
18. В … в явном виде указаны семантические связи между лексическими единицами
а) энциклопедии; б) словоуказателе; в) конкордансе; г) *тезаурусе*

19. Главное слово тезауруса называется …
а) *дескриптор*; б) прескриптор; в) транскриптор; г) ключевое слово

Тестовые задания по теме: «Основные понятия электронной лексикографии. Лексикографические параметры электронных словарей разных типов**»**
1. В рамках компьютерной лексикографии разрабатываются …
а) *компьютерные технологии составления и использования словарей;*
б) базы данных, способные обрабатывать информационные массивы;
в) компьютерная словарная терминология;
г) эксперименты и алгоритмы построения терминологических словарей.
2. Автоматические словари отличаются по …
а) своему назначению;
б) *интерфейсу*;
в) своей сложности;
г) своей цене.
3. Особенности структуры автоматических словарей зависят от …
а) цены;
б) языка, на котором он написаны;
в) *тех программ, с которыми они взаимодействуют;*
г) пользователей, для которых они предназначены.
4. Компьютерная лексикография, занимающаяся представлением информации, лишена…
а) *собственного предмета исследования*;
б) финансовой поддержки исследований;
в) недостатков бумажных словарей;
г) возможности успешно развиваться.
5. Функция компьютерной лексикографии…
а) развитие бумажной лексикографии;
б) предоставлением современных возможностей обработки информации лексикографом;
в) *удобное представление уже существующего содержания;*
г) сделать словари доступными по цене населению.
6. Какие из задач не стоят перед составителями электронных словарей?
а) добавление различных опций работы с лексикографическими материалами;
б) использование новейших лексикографических концепций, которые становятся возможными именно благодаря компьютерной лексикографии;
в) снижение себестоимости издания бумажных словарей;
г) развитие форматирования словарных статей.
7. Проблемы бумажных словарей:
а) чем больше объем словаря, тем сложнее им пользоваться;
б) отсутствует возможность быстрого редактирования;
в) высокая цена;
г) опасность аллергических реакций человека на пылевого клеща.
8. Какие возможности не относятся к компьютерной лексикографии?
а) возможность использования большего объема лексикографической информации;
б) использование различных программ для облегчения труда лексикографа;
в) более широкие возможности содержания словарной статьи;
г) возможность показа по критериям;
д) использование различных лингвистических технологий (морфологический, синтаксический анализ, полнотекстовый поиск, распознавание/синтез звука, генерация текста;
е) развитие рынка сбыта посредством размещение лексикографических источников в сети Интернет.
9. В современных электронных переводных словарях единицей описания становится …

а) термин;
б) калька;
в) слово или словосочетание;
г) предложение.
10. Назовите, что не является программами поддержки лексикографических работ …
а) компьютерные картотеки;
б) программы обработки текста;
в) программы, помогающие формировать статьи в автоматическом режиме;
г) программы разработки интерфейса словаря.

Тестовые задания по теме: «Корпусная лингвистика: предмет, терминология. Связь корпусной лингвистики с др. науками. Методы корпусной лингвистики**»**

1. Как Вы понимаете термин «корпусная лингвистика»?
а) *создание и использование электронных корпусов текстов;*
б) создание электронных словарей, тезаурусов, онтологий. Например, Lingvo. Словари используют, например, для автоматического перевода, проверки орфографии;
в) автоматический перевод текстов. Среди русских переводчиков популярным является Промт. Среди бесплатных известен переводчик Google Translate;
г) создание вопросно-ответных систем (англ. *question answering systems*).

2. По Э. Финегану корпус – это …
а) большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач;
б) *репрезентативное собрание текстов, обычно в машиночитаемом формате и включающее информацию о ситуации, в которой текст был произведен, такую как информация о говорящем, авторе, адресате или аудитории;*
в) собрание языковых фрагментов, отобранных в соответствии с четкими языковыми критериями для использования в качестве модели языка;
г) собрание текстов, в основе которого лежит логический замысел, логическая идея, объединяющая эти тексты и воплощенная в правилах организации текстов в корпус, алгоритме и программе анализа корпуса текстов, сопряженной с этим идеологии и методологии.

3. Корпусный менеджер – это…
а) операционщик, обслуживающий сайт, на котором размещен корпусный ресурс;
б) администратор, занятый управлением, программированием и обновлением корпусного сайта;
в) система планирования работ, просмотр отчетов и контроль выполнения задач корпусного ресурса;
г) *система управления текстовыми и лингвистическими данными.*

4. Поиск в корпусе позволяет по любому слову построить конкорданс – это …
а) архив, который используется для объединения множества любых файлов в единый файл-контейнер с целью удобства хранения и переноса информации;
б) организованная структура, предназначенная для хранения, изменения и обработки взаимосвязанной информации корпуса;
в) *список всех употреблений данного слова в контексте со ссылками на источник;*
г) онтология, которая представляет собой множество классов, связанных между собой отношением обобщения лингвистических данных.

5. В каком из определений подчеркивается созидательная направленность корпусной лингвистики?
а) деятельность, связанная с программированием языковых ресурсов;
б) деятельность, обусловленная и продиктованная современной действительностью;

в) *деятельность, требующаяся для составления и использования корпуса, направленная на исследование естественного употребления языка;*
г) деятельность, выявляющая структуры языка в их взаимодействии.

6.Корпусная лингвистика имеет своим предметом…
а) моделирование процесса понимания смысла текстов (перехода от текста к формализованному представлению его смысла) и проблема синтеза речи (перехода от формализованного представления смысла к текстам на естественном языке);
б) *теоретические основы и практические механизмы создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей;*
в) изучение и разработка способов оптимизации различных сфер функционирования языковой системы;
г) разработка и изучение понятий, образующих основу формального аппарата для описания строения естественных языков (т. е. метаязыка лингвистики).

7. Терминология корпусной лингвистики в русском языке…
а) *в процессе становления;*
б) не сложилась;
в) сложилась;
г) затрудняюсь ответить.

8. Отметьте варианты терминов, которые употребляются по отношению к корпусной лингвистике:
а) *корпусы;*
б) корпуса;
в) *ко́рпусный*;
г) корпусно́й.

9. Технологии, которые применяются в корпусной лингвистике, намного старше…
а) библиотечных картотек;
б) *электронных компьютеров*;
в) глоссариев;
г) систем учета информации.

10. Лингвисты собрали первые корпусы компьютеризированных текстов в ….
а) *1960-е гг.;*
б) 1970-е гг.;
в) 1980-е гг.;
г) 1990-е гг.

11. Первым компьютеризированным корпусом был…
а) Ланкастер-Осло-Берген корпус (The Lancaster-Oslo-Bergen Corpus);
б) Британский национальный корпус (British National Corpus);
в) Корпус современного американского английского (Corpus of Contemporary American English);
г) *Брауновский корпус (The Brown Corpus).*

12. Когда корпусная лингвистика окончательно сформировалась как отдельное направление науки о языке?
а) в первой половине 1980-х годов
б) *в первой половине 1990-х годов;*
в) в первой половине 2000-х годов;
г) в первой половине 2010-х годов.

13. Репрезентативность, или сбалансированность корпуса – это...
а) склонность чувствовать себя преуспевающим, имеющим приличный вид;
б) соразмеренность операционных действий в корпусном ресурсе;
в) *представительность корпуса, соотношение его отдельных частей (по разным характеристикам);*

г) уравновешенность методик при использовании данных корпуса.

14. Задача авторов корпуса состоит в том, чтобы…

а) установить типологию входящих в корпус текстов;

б) определить причины и относительную хронологию включения в корпус текстов;

в) исследовать процессы языковых изменений языка корпуса;

г) *собрать как можно большее количество текстов, относящихся к тому подмножеству языка, для изучения которого корпус создается.*

15. Отобранные тексты в Брауновском корпусе должны были отражать…

а) *15 регистров;*

б) 16 регистров;

в) 17 регистров;

г) 18 регистров.

16. Перечислите некоторые из регистров, которые были положены в основу Брауновского корпуса.

а) _____;

б) _____;

в) _____;

г) _____.

17. Методология построения корпусов первого типа основывается на…

а) применении модели как средства исследования;

б) корректно отраженных частных, единичных лингвистических феноменах в корпусе текстов, специально созданном для их отражения;

в) типе корпуса;

г) *реализации проблемы корректности движения от объективно существующей речевой практики носителей языка к частному корпусу текстов.*

18. Методология построения корпусов второго типа должна…

а) реализовать проблемы корректности движения от объективно существующей речевой практики носителей языка к частному корпусу текстов;

б) применять модели как средства исследования;

в) *корректно отражать частные, единичные лингвистические феномены в корпусе текстов, специально созданном для их отражения;*

г) основываться на принципах двойной записи, балансе и теории равновесия.

19. Приведите примеры типов корпусов…

а) _____;

б) _____;

в) _____;

г) _____.

20. Главная сложность создания фонетических лингвистических ресурсов связана с…

а) необходимостью осуществлять диктофонную цифровую запись речи;

б) необходимостью находить дикторов, обладающих нормированной речью;

в) *необходимостью транскрибирования устной речи;*

г) необходимостью аннотировать записанные речевые образцы.

Тестовые задания по теме: «Способы использования корпусов. Типы корпусов**»**

1. Корпус является продолжением…

а) интернетного сайта;

б) компьютерной программы;

в) *традиционных картотек;*

г) лингвистического исследования.

2. Репрезентативность корпуса должна обеспечиваться…

а) достоверностью данных;

б) количеством текстовой выборки;

в) оперативностью поиска искомых единиц;
г) *достаточным объемом текстового материала, так и его разнообразием.*

3. Какие из этапов создания корпусов являются излишними, необязательными?
а) обеспечение поступления текстов в соответствии с перечнем источников;
б) *составление традиционной картотеки данных;*
в) преобразование её в машиночитаемую форму;
г) разметка текста.

4. Что обеспечивает конвертирование размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager)?
а) *быстрый многоаспектный поиск и статистическую обработку;*
б) управление ресурсами корпусного сайта;
в) программирование корпусного ресурса;
г) отладка системы и структуры корпуса.

5. Какой метод в сочетании с опытом специалистов был использован при создании корпуса текстов «Американский корпус наследия» (The American Heritage Intermediate Corpus)?
а) метод интервью;
б) статистический метод;
в) *метод анкет;*
г) метод записи телефонных разговоров.

6. Какая процедура обработки письменного языка является необязательной?
а) токенизация;
б) лемматизация;
в) парсинг;
г) *вебинг.*

7. Что такое токенизация?
а) *разбиение потока символов в естественном языке на отдельные значимые единицы (токены, словоформы);*
б) систематизация токенов по определенным критериям;
в) создание токенов в процессе сегментирования письменных текстов;
г) программирование особых токенов для классифицирования классов и подгрупп.

8. Парсинг – это процесс сопоставления линейной последовательности лексем (слов, токенов) языка с его …
а) формальной фонетикой;
б) *формальной грамматикой;*
в) формальной лексикой;
г) формальной стилистикой.

9. Для каких видов анализа применяются такие программные средства как тэггеры (taggers) и парсеры (parsers)?
а) *морфологического*;
б) анафорического;
в) *синтаксического*;
г) интонационного.

10. Почему автоматический анализ естественного языка небезошибочен?
а) всегда имеют место ошибки;
б) после такого вида анализа нужно делать корректировку вручную;
в) имеет место несовершенство компьютерных систем;
г) *такой анализ дает несколько вариантов анализа для одной лексической единицы (слова, словосочетания, предложения).*

11. Для решения различных лингвистических задач недостаточно иметь массив текстов. Требуется также..

а) автоматическая разметка корпусных текстов;
б) *чтобы тексты содержали в себе явным образом указанную разного рода дополнительную лингвистическую и экстралингвистическую информацию;*
в) транскрипция для изучения звуковой оболочки слов;
г) помощь разработчиков корпусного ресурса.

12. Что такое «проблема морфологической неоднозначности (ambiguity)»?
а) имеется неоднозначный морфологический разбор слова по составу;
б) программные средства с точностью не могут определить словообразовательную форму слова;
в) *некоторые формы слов могут быть членами более чем одной грамматической категории;*
г) морфологическая неоднозначность перекрещивается с синтаксической неоднозначностью.

13. Разметка заключается в приписывании текстам и их компонентам специальных тэгов (укажите каких и расшифруйте каждый тег):
а) _____
_____;
б) _____
_____.

14. Синтаксическая разметка является результатом …, выполняемого на основе данных морфологического анализа
а) токенизации;
б) лемматизации;
в) *парсинга;*
г) вебинга.

15. Синтаксическая разметка описывает синтаксические связи между … единицами и различные синтаксические конструкции
а) *лексическими*;
б) звуковыми;
в) морфологическими;
г) синтаксическими.

16. Сколько групп Е-факторов, влияющих на язык текстов, выделяет Дж. Синклер?
а) 2;
б) *3*;
в) 4;
г) 5.

17. Зачем необходимы единые форматы представления данных разных корпусов?
а) *это позволяет во многих случаях использовать единое программное обеспечение и обмениваться корпусными данными;*
б) это требование международных стандартов при создании корпусов;
в) это уменьшает временные и человеческие затраты на создание корпусных ресурсов;
г) это сокращает финансовые издержки на программирование сайтов.

18. В качестве формального языка разметки текстов в корпусе широко применяются языки …
а) Javascript и Python;
б) Ruby и PHP;
в) *SGML и XML;*
г) C++ и Objective-C.

19. Одним из наиболее эффективных корпусных менеджеров является …
а) SARA;
б) XAIRA (BNC);
в) CQP;

г) *Bonito/Manatee*.

20. При использовании веб-пространства как корпуса роль корпусных менеджеров выполняют

а) автоматические регуляторы;
б) закодированные программы;
в) цифровые операторы;
г) *поисковые системы*.

2.1.4. ПРАКТИКО-ОРИЕНТИРОВАННЫЕ ЗАДАНИЯ (КЕЙС-ЗАДАЧИ)

Задание 1. Проанализируйте систему согласных фонем эвенкийского языка. Подберите в Речевом корпусе эвенкийского языка (аннотированном) примеры на реализацию каждой из согласных фонем в различных комбинаторно-позиционных условиях. Составьте матрицу аллофонов согласных фонем в текстовом редакторе. Дайте пример на каждый аллофон, затранскрибируйте примеры с помощью системы МФА, используя шрифт Doulous SIL.

Задание 2. В системах электронных библиотек найдите информацию об акустических характеристиках и реализации сонорных в разных языках мира. В аудиоархиве Лаборатории экспериментально-фонетических исследований найдите эвенкийские звуковые образцы, содержащие губно-губной звонкий сонорный /β/. Составьте список слов, содержащих данный звук. Выберите из составленного списка 10 словоформ и проаннотируйте их в программе по обработке звукового сигнала PRAAT. Аннотация должна состоять из 6 слоев: word, intonation, phoneme, sillable, allophone, bell. Транскрибирование звуковых сегментов обозначать средствами МФА.

Задание 3. Используя литературные источники, проанализируйте систему гласных фонем эвенкийского языка. Методом сплошной выборки осуществите поиск реализаций гласных фонем на основе корпусного ресурса «Речевой корпус эвенкийского языка (аннотированный)». Количество реализаций каждой эвенкийской фонемы в Вашем списке не должно быть меньше 20 единиц. В компьютерной программе по обработке звукового сигнала PRAAT произведите замеры частотных характеристик каждой из гласных фонем. Результаты замеров внесите в таблицу «Частотные характеристики гласных эвенкийского языка». На основе таблицы постройте точечную диаграмму, показывающую вокалический трапецоид, свойственный артикуляции выбранного диктора.

Задание 4. При помощи электронных библиотечных систем найдите информацию о языках мира, в которых имеется противопоставление гласных фонем по оппозиции «долгота-краткость». Изучая этот вопрос, обратите внимание на реализацию этого признака «фонологическая долгота гласных» в тюркских (бурятском, якутском) и тунгусо-маньчжурских языках. Методом сплошной выборки осуществите поиск реализаций долгих и кратких гласных фонем на основе корпусного ресурса «Речевой корпус эвенкийского языка (аннотированный)». Количество реализаций каждой эвенкийской фонемы в Вашем списке не должно быть меньше 20 единиц. В компьютерной программе по обработке звукового сигнала PRAAT произведите замеры количественных характеристик каждой из гласных фонем. Результаты замеров внесите в таблицу «Количественные характеристики гласных эвенкийского языка». На основе таблицы постройте диаграмму.

Задание 5. Повторите информацию о таких сложных в артикуляционном смысле звуках как аффрикаты. На основе «Речевого корпуса эвенкийского языка» произведите поиск 150 словоформ, содержащих среднеязычный глухой аффрицированный согласный /ʧ/.

Произведите замеры длительности фаз аффрикат: смычки, взрыва, щелевой фазы. Результаты замер внесите в таблицу «Реализация аффрикат в речи диктора N». По результатам таблицы постройте диаграмму.

Задание 6. Повторите информацию об аллофоном варьировании гласных в различных комбинаторно-позиционных условиях. На основе «Речевого корпуса эвенкийского языка» произведите поиск 20 словоформ на каждый гласный в речи одного из дикторов-носителей селемджинского говора эвенкийского языка. Проанализируйте комбинаторно-позиционные условия, в которых реализовался каждый гласный, осуществляя замеры частотных характеристик гласных сегментов в программе по обработке речевого сигнала PRAAT. Данные замеров занесите в таблицу «Частотные характеристики гласных в различных комбинаторно-позиционных условиях». Опишите таблицы.

Задание 7. В системах электронных библиотек найдите информацию об акустических характеристиках и реализации сонорных в разных языках мира. В аудиоархиве Лаборатории экспериментально-фонетических исследований найдите эвенкийские звуковые образцы, содержащие губно-губной звонкий сонорный /l/. Составьте список слов, содержащих данный звук. Выберите из составленного списка 10 словоформ и проаннотируйте их в программе по обработке звукового сигнала PRAAT. Аннотация должна состоять из 6 слоев: word, intonation, phoneme, sillable, allophone, bell. Транскрибирование звуковых сегментов обозначать средствами МФА.

Задание 8. Из аудиоархива по эвенкийскому языку выберите нераспространенные предложения, соответствующие таким коммуникативным типам высказываний как утверждение, общий вопрос, специальный вопрос, альтернативный вопрос, восклицание, побудительное высказывание. В программе по обработке звукового сигнала PRAAT создайте интонограмму для каждого коммуникативного типа. Проанализируйте каждую из полученных интонограмм. Результаты замеров частоты основного тона (ЧОТ), скорости инклинации и диклинации, частотные характеристики интонационного центра занесите в таблицу.

Задание 9. Изучите литературу, посвященную описанию реализации коррелятов ударения в агглютинативных языках. На материале двусложных слов «Речевого корпуса эвенкийского языка (аннотированного)» из речи одного из дикторов селемджинского говора проведите акустический анализ в программе PRAAT тонового и количественного коррелятов словесного ударения в эвенкийском языке. Результаты замеров частоты основного тона (ЧОТ) оформите в таблицу.

2.1.5. ПРЕЗЕНТАЦИЯ

Электронная презентация – это электронный документ, позволяющий последовательно показывать в полноэкранном режиме слайды. Слайд – основной элемент презентации. Слайды могут содержать мультимедийную информацию разных типов: текстовую, графическую (диаграммы, графики, картинки), фотографии, видеофрагменты и аудиоинформацию (звуковое сопровождение, закадровый текст). Электронная презентация предназначена для демонстрации объектов и событий, которые не могут быть непосредственно представлены аудитории, во время выступления докладчика. В зависимости от назначения презентация может быть выполнена в любых программных продуктах, позволяющих отображать слайды с заданным промежутком времени или по требованию докладчика. Информация, представленная с помощью мультимедийной презентации, быстро усваивается за счет сокращения текстовой части и замены словесных описаний объекта яркими иллюстрациями, наглядными схемами и графиками.
Требования, позволяющие создавать наиболее эффективные электронные презентации:

1. **Содержание мультимедийной презентации**: отображение темы электронной презентации и данных об авторе презентации или учреждении на первом слайде презентации; соответствие содержания презентации поставленным целям и задачам; краткость изложения, максимальная информативность и достоверность представленной информации; новизна и конкурентоспособность презентуемого объекта; формулировка запоминающейся ключевой фразы презентации.
2. **Расположение информации на слайде**: горизонтальное расположение информации на слайде, форматирование текста по ширине, размещение наиболее важной информации в центре экрана и выделение ее шрифтом и цветом; вставка надписей под мультимедийной информацией.
3. **Визуальный и звуковой ряд**: соответствие изображений и графических объектов содержанию презентации; обеспечение яркости и контрастности изображения; обеспечение высокого качества используемой аудиоинформации.
4. **Текст**: использование контрастных цветов для фона и текста; выделение ключевых слов в предложении жирным шрифтом или цветом; соблюдение принятых правил орфографии, пунктуации, сокращений и правил оформления текста (отсутствие точки в заголовках и др.); недопустимость применения переносов в словах; использование подчеркивания лишь в гиперссылках.
5. **Дизайн**: использование единого стиля оформления презентации; соответствие стиля оформления (графического, звукового, анимационного) содержанию презентации; использование психологически комфортного фона слайдов, не заслоняющего информацию, представленную на них; соответствие шаблона представляемой теме; целесообразность использования анимационных эффектов.
6. **Адаптивность мультимедийной презентации**, возможность внесения в нее изменений и дополнений в зависимости от конкретной ситуации .
7. **Время**: количество слайдов примерно соответствует длине доклада в минутах (если у вас слайдов намного больше, чем времени, вы не успеете показать все слайды, либо будете показывать их слишком быстро и аудитория не поймет доклада; если у вас слайдов слишком мало, то это означает, что вы не эффективно их используете).
8. **Доклад**: Повествование должно быть последовательным и логичным.• Возврат к старому слайду – зачастую неудачное решение.• Доклад следует делить на разделы.• Если доклад длится более 15-20 минут, следует перед каждым разделом давать его краткий• план. В каждом разделе не должно быть более 4-5 параграфов (иначе к тому времени, как вы• дойдете до последнего параграфа, аудитория успеет забыть план этого раздела). Названия разделов и параграфов должны быть краткими и ёмкими.• Начинайте доклад с пояснения, о чем вы будете рассказывать.• Завершайте свой доклад обобщением уже сказанных основных тезисов в более короткой и• понятной форме. Люди наиболее внимательны в начале и конце доклада. Итоги – это второй шанс донести главную мысль до слушателя.

Задание: Выполните проект на тему «Разработка лингвистических компонентов «Речевого корпуса эвенкийского языка (аннотированного)», Представьте результаты проекта в виде презентации.

Форма представления задания: доклад выступления, презентация.

Цель работы (проекта): формирование профессиональных компетенций дисциплины «Общая и компьютерная лексикография и технологии корпусной лингвистики»:

владением основными понятиями и категориями современной лингвистики (ОПК-1);

способностью пользоваться лингвистически ориентированными программными продуктами (ПК-9);

владением принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологий; фонетических, лексических, грамматических и иных баз и баз знаний) и умением пользоваться такими ресурсами (ПК-10);

способностью использовать лингвистические технологии для проектирования систем автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистических компонентов интеллектуальных и информационных систем (ПК-11).

Задачи работы:

- развитие навыков исследовательской работы по анализу языка на базе корпусных данных;
- применение современных технологий сбора, обработки и интерпретации полученных экспериментальных данных;
- совершенствовать навыки работы с компьютерными программами обработки речевого сигнала «Audacity»;
- совершенствовать навыки публичного выступления с опорой на визуальные средства представления информации.

В ходе проекта следуйте алгоритму действий:

1) проведите мониторинг информационных массивов в сети Интернет по речевым корпусам и базам данных эвенкийского языка и подготовьте на основе мониторинга аналитические материалы;
2) составьте список из десяти звуковых образцов речи эвенков Приамурья, содержащие палатализованные аллофоны шумных согласных, используя аудиоархив Лаборатории экспериментально-фонетических исследований кафедры иностранных языков АмГУ;
3) нормализуйте в компьютерной программе по обработке звукового сигнала «Audacity» звуковые образцы, отобранные в аудиоархиве ЛЭФИ, удалите шумы. Если необходимо, выполните усиление уровня громкости. Экспортируйте звуковые образцы в формат .wav
4) добавьте созданные лингвистические компоненты на кафедральный сайт «Речевого корпуса эвенкийского языка (аннотированного)» https://linguacorpus.amursu.ru/
5) результаты проекта изложите в докладе и презентации.

Разместите результаты работы в портфолио:

а) текст научного доклада;

б) презентация.

Примерные темы для подготовки презентаций по темам дисциплины

1. Лексикография как наука. Словарь как способ семантизации общей и специальной лексики в прикладных целях.
2. Типология словарей. Типологические параметры словаря. Современные классификации словарей.
3. Формализация структуры словаря. Лингвистическое и компьютерное обеспечение словарей.
4. Типы информации в словаре и компьютерной базе данных.
5. Типы терминологических словарей
6. Основные понятия электронной лексикографии.
7. Элементы систем управления базами данных. Таблицы, формы, фильтры, запросы, отчеты.
8. Идеографическая лексикография.
9. Предмет корпусной лингвистики
10. Полевая лингвистика и корпусная лингвистика
11. Связь корпусной лингвистики с этнолингвистикой
12. История лингвистических корпусов: от картотеки к корпусу.
13. Классификация (типология) корпусов.

14. Корпусная лингвистика: современное состояние.
15. Корпусная лингвистика в России.
16. Обзор существующих корпусов различных типов.
17. Корпус как поисковая система.
18. Корпусоподобные интерфейсы между лингвистом и поисковыми системами Интернета.
19. Лингвистические исследования, базирующиеся на корпусах.
20. Проблемы репрезентативности корпусов.
21. Проблемы хронологии в общеязыковых корпусах.
22. Отбор текстов для корпусов.
23. Графематический анализ.
24. Понятие разметки.
25. Типы разметки.
26. Морфологическая разметка.
27. Синтаксические корпуса (treebanks).
28. Семантическая разметка.
29. Технология создания корпусов. Стадии работы.
30. Понятие корпусоида.
31. Автоматическая морфоразметка.
32. Автоматический синтаксический анализ (parsing).
33. Языковые средства представления размеченных текстов (языки SGML, XML).
34. Международные стандарты (TEI, EAGLES, CDIF, XCES).
35. Способы использования корпусов в лингвистических исследованиях.
36. Исследование способов использования корпусов в лексикографии.
37. Изучение средств обработки корпусных данных, представленных на языке XML.
38. Обзор существующих корпусов различных типов.
39. Классификация (типология) корпусов по различным основаниям.
40. Типы корпусов по задачам. Типы корпусов по формальным признакам.

2.2. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО САМОСТОЯТЕЛЬНОЙ РАБОТЕ СТУДЕНТОВ

Написание терминологических диктантов и тестов, выполнение практических заданий, подготовка докладов по темам дисциплины, которое сопровождается слушанием лекций и чтением учебной и научной литературы, способствует развитию самостоятельного лингвистического мышления и выработке научно-исследовательского подхода языковому материалу.

При подготовке к терминологическим диктантам необходимо пользоваться современными признанными лингвистической наукой лингвистическими словарями и справочниками. Поскольку лингвистическая терминология является открытой, регулярно пополняющейся системой желательно оказывать предпочтение мотивированным терминам, имеющим прозрачную смысловую структуру.

При подготовке к лабораторным занятиям каждый студент должен:

- изучить рекомендованную учебную литературу;
- изучить конспекты лекций;
- подготовить ответы на все вопросы по изучаемой теме.

По согласованию с преподавателем студент может подготовить доклад или сообщение по теме лабораторного занятия. В процессе подготовки к лабораторным занятиям студенты могут воспользоваться консультациями преподавателя.

Вопросы, не рассмотренные на лекциях и лабораторных занятиях, должны быть изучены студентами самостоятельно. Контроль самостоятельной работы студентов по учебной программе курса осуществляется в ходе лабораторных занятий методом устного опроса, письменных заданий или посредством тестирования. В ходе самостоятельной работы студент обязан прочитать рекомендуемую основную и дополнительную

литературу по изучаемой теме, дополнить конспекты лекций недостающим материалом, выписками из рекомендованных источников. Выделить непонятные термины, найти их значение в словарях. Студент должен готовиться к предстоящему практическому занятию по всем обозначенным в программе вопросам. Вызвавшие у студента в ходе самостоятельной работы затруднение вопросы следует прояснить на семинарских занятиях.

При изучении дисциплины используются следующие виды самостоятельной работы студентов:

- поиск (подбор) литературы (в том числе электронных источников информации) по заданной теме, сравнительный анализ научных публикаций;
- подготовка докладов и сообщений по вопросам курса для лабораторных занятий;
- работа над терминологией курса.

Для подготовки к занятиям, текущему контролю и промежуточной аттестации студенты могут воспользоваться электронными библиотеками, а также могут взять на дом необходимую литературу на абонементе вузовской библиотеки и воспользоваться читальными залами вуза.

Рекомендации по планированию и организации времени, необходимого на изучение дисциплины

Наиболее оптимальный вариант планирования и организации студентом времени, необходимого для изучения дисциплины, – распределить учебную нагрузку равномерно, т. е. каждую неделю знакомиться с необходимым теоретическим материалом на лекционных занятиях и закреплять полученные знания самостоятельно, прочитывая рекомендуемую литературу.

К занятиям необходимо готовиться за неделю или две до срока их проведения, чтобы была возможность проконсультироваться с преподавателем по трудным вопросам. В случае пропуска занятия, необходимо предоставить письменную разработку пропущенной темы. Самостоятельную работу следует выполнять согласно графику и требованиям, предложенным преподавателем.

Допуск к экзамену по дисциплине предполагает активное участие на занятиях, а также своевременное выполнение домашних и самостоятельных заданий.

Описание последовательности действий студента при изучении дисциплины

Задание для подготовки к занятиям по данному курсу студент получает от преподавателя.

Основным промежуточным показателем успешности студента в процессе изучения дисциплины является его готовность к занятиям.

Приступая к выполнению задания по любой теме, прежде всего, необходимо

- ознакомиться с планом занятия,
- изучить соответствующий раздел учебного пособия,
- выяснить наличие литературы или теоретического материала по соответствующей теме,
- по каждому вопросу предложенной темы необходимо определить и усвоить ключевые понятия и термины,
- для более глубокого понимания проблемы необходимо познакомиться с дополнительной литературой и законспектировать основные положения.

В случае возникновения трудностей студент должен и может обратиться за консультацией к преподавателю.

Критерием готовности к занятию является умение ответить на все вопросы по теме занятия, а также наличие соответствующих конспектов.

Рекомендации по подготовке к экзамену

В процессе подготовки к экзамену рекомендуется:

1) ознакомиться с перечнем вопросов, выносимых на экзамен;

2) повторить, обобщить и систематизировать информацию, полученную на протяжении всего учебного года в процессе слушания лекций, чтения учебников, учебных пособий, монографий, сборников научных статей, журналов и газетных публикаций, предлагаемых для углубленного изучения той или иной темы;
3) просмотреть конспекты лекций, карты-конспекты, содержащие основные положения концепций авторов, работы которых изучались во время самостоятельной работы;
4) выучить определения основных понятий и категорий, повторить терминосистему изученного курса.

Рекомендации по работе с литературой. При работе с литературой необходимо, во-первых, определить, с какой целью студент обращается к источникам: найти новую, неизвестную информацию; расширить, углубить, дополнить имеющиеся сведения; познакомиться с другими точками зрения по определенному вопросу; научиться применять полученные знания; усовершенствовать умения. Исходя из этих целей, необходимо выбрать источники. Прежде всего, следует обратиться к учебникам, названия которых совпадают с названием курса. Для формирования умений целесообразно обратиться к практикумам. В получении более глубоких знаний по отдельным темам, проблемам помогут научные статьи, монографии, книги, приведенные в списках дополнительной литературы. При подготовке докладов и сообщений целесообразно обратиться также к научно-популярной литературе.

Выбрав несколько источников для ознакомления, необходимо изучить их оглавление. Это позволит определить, представлен ли там интересующий вопрос, и в каком объеме он освещается. После этого откройте нужный раздел, параграф и просмотрите, пролистайте их, обратив внимание на заголовки и шрифтовые выделения, чтобы выяснить, как изложен необходимый материал в данном источнике (проблемно, доступно, очень просто, популярно интересно, с представлением разных позиций, с примерами и проч.). Так, на основании ознакомительного, просмотрового чтения из нескольких книг, статей вы выберете одну-две (для подготовки доклада больше) для детальной проработки.

После этого переходите к изучающему и критическому видам чтения: фиксируйте в форме тезисов, выписок, конспекта основные, значимые положения, отмечайте свое согласие с автором или возможные спорные моменты, возражения. При этом известную информацию вы пропускаете, ищите в данном источнике новое, дополняющее ваши знания по предмету, определяя, что из этого важно, а что носит факультативный, дополнительный, может быть занимательный характер. Обязательно укажите авторов, название, выходные данные источника, с которым вы работали, т. е. оформите библиографические сведения о нем.