Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Северо-Осетинский государственный университет имени Коста Левановича Хетагурова»

Кафедра осетинского языка

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ДИСЦИПЛИНЕ «Корпусные исследования в современной лингвистике»

Направление подготовки 45.04.01 Филология
программа: «Языки народов Российской Федерации (осетинский язык)»

Составители: старший преподаватель кафедры прикладной математики Ф.Х. Мамсурова; профессор кафедры осетинского языка и литературы, доктор филологических наук Л.Б. Гацалова.

.

Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Методические рекомендации по дисциплине

Вид работ	Методические рекомендации
лекции	В ходе лекционных занятий вести конспектирование учебного материала. Обращать внимание на категории, формулировки, раскрывающие суть тех или иных явлений и процессов, научные выводы и практические рекомендации, положительный опыт в ораторском искусстве. Желательно оставить в рабочих конспектах поля для пометок.
практические занятия	Работа на практических занятиях предполагает активное участие в дискуссиях. Важной формой самостоятельной работы обучающегося является систематическая и планомерная подготовка к практическому занятию. После лекции следует познакомиться с планом практических занятий и списком обязательной и дополнительной литературы, которую необходимо прочитать, изучить и законспектировать. Разъяснение по вопросам новой темы даются преподавателем в конце предыдущего практического занятия.
самостоятельная работа	САМОСТОЯТЕЛЬНАЯ РАБОТА требует, прежде всего, чтения рекомендуемых источников и монографических работ, их реферирования, подготовки докладов и сообщений. Важным этапом в самостоятельной работе является повторение материала по конспекту лекции. Одна из главных составляющих внеаудиторной подготовки - работа с книгой. Она предполагает: внимательное прочтение, критическое осмысление содержания, обоснование собственной позиции по дискуссионным моментам, постановки интересующих вопросов, которые могут стать предметом обсуждения на семинаре. При работе с терминами необходимо обращаться к словарям, в том числе доступным в Интернете, например, на сайте http://dic.academic.ru.
доклад	Доклад - краткое изложение в письменном виде полученных результатов теоретического анализа определѐнной научной (учебно-исследовательской) темы, в рамках которой автор раскрывает суть исследуемой проблемы, приводит различные точки зрения, а также собственные взгляды на неѐ. Тема доклада (его объем - от 10 до 15 машинописных страниц без учета приложений) соответствует одному из вопросов, номер которого совпадает с последней цифрой номера студента в списке группы. На основе реферативного обзора готовится выступление по рассматриваемой проблеме на 5-7 минут. Структура доклада включает в себя: титульный лист, содержание, введение, разделы основной части, заключение, список использованных источников и возможно приложения. Текст доклада необходимого набирать на компьютере на одной стороне листа. Размер левого поля 20 мм, правого - 10мм, верхнего - 20мм нижнего - 20мм. Шрифт Times New Roman, размер - 14, межстрочный интервал - 1,5. Фразы, начинающиеся на с новой строки, печатаются с абзацным отступом от начала строки. Доклад, выполненный небрежно, неразборчиво, без соблюдения требований по оформлению, возвращается студенту без проверки с указанием причин возврата на титульном листе.
реферат	Реферат – продукт самостоятельной работы аспиранта, представляющий собой краткое изложение в письменном виде полученных результатов теоретического анализа определенной научной (учебно-исследовательской) темы, где автор раскрывает суть исследуемой проблемы, приводит различные точки зрения. В РПД приводится перечень тем, среди которых аспирант может выбрать тему реферата. С защитой своего реферата аспирант выступает на семинарском занятии (время выступления – 10 мин.). При

	оценке реферата (собственно текста и процедуры защиты) критериями выступают: − информационная достаточность; − соответствие материала теме и плану; − стиль и язык изложения (целесообразное использование терминологии, пояснение новых понятий, лаконичность, логичность, правильность применения и оформления цитат и др.); − наличие выраженной собственной позиции; − адекватность и количество использованных источников (7– 10); − владение материалом.
конспект	Конспект позволяет формировать и оценивать умения аспирантов по переработке информации. При оценке конспекта критериями выступают: – оптимальный объем текста (не более одной трети оригинала); – логическое построение и связность текста; – полнота/ глубина изложения материала (наличие ключевых положений, мыслей); – визуализация информации как результат её обработки (таблицы, схемы, рисунки); – оформление (аккуратность, соблюдение структуры оригинала).
презентация	1) Не перегружать слайды текстом. 2) Наиболее важный материал лучше выделить. 3) Не следует использовать много мультимедийных эффектов анимации. Особенно нежелательны такие эффекты, как вылет, вращение, побуквенное появление текста. Оптимальная настройка эффектов анимации – появление, в первую очередь, заголовка слайда, а затем текста по абзацам. При этом если несколько слайдов имеют одинаковое название, то заголовок слайда должен постоянно оставаться на экране. 4) Чтобы обеспечить хорошую читаемость презентации необходимо подобрать темный цвет фона и светлый цвет шрифта. 5) Текст презентации должен быть написан без орфографических и пунктуационных ошибок.
Собеседование	Собеседование – средство контроля, организованное как специальная беседа преподавателя с обучающимся на темы, связанные с изучаемой дисциплиной и рассчитанное на выяснение объема знаний аспиранта по определенному вопросу (из перечня вопросов к зачету. При оценивании результатов собеседования критериями оценки результатов выступают: − усвоения знаний (глубина, прочность, систематичность знаний); − умений применять знания (адекватность применяемых знаний в конкретной ситуации); − рациональность используемых подходов, умение логически выстроить ответ; − сформированность профессионально значимых личностных качеств; − коммуникативные навыки (умение поддерживать и активизировать беседу).
контрольная работа	Контрольная работа - письменная работа, выполняемая по дисциплине, в рамках которой раскрываются конкретные темы с целью оценки качества усвоения студентами отдельных, наиболее важных разделов, тем и проблем изучаемой дисциплины. Оценить умение обучающегося письменно излагать материал по конкретной теме, аргументировано и структурировано излагать суть поставленной проблемы, анализировать представленные позиции, делать выводы и уметь представить собственную позицию по поставленной проблеме. Студенты заочной формы обучения в соответствии с учебным планом и программой выполняют по курсу дисциплины одну контрольную работу. Контрольная работа включает один теоретический вопрос. Вариант зада-

	ния на контрольную работу определяется преподавателем. Выполняя контрольную работу, необходимо показать умение правильно, коротко и четко излагать усвоенный материал. В процессе подготовки к выполнению контрольной работы следует изучить рекомендованную литературу, а также новые публикации в области дисциплины в периодической печати. При написании ответов на вопросы желательно приводить цитаты, которые должны иметь ссылки на информационный источник (фамилия, инициалы автора, название цитируемого источника, том, часть, выпуск, издательство, год, страница). При выполнении контрольной работы следует творчески подходить к имеющейся информации, уметь выразить свое мнение по исследуемому вопросу. Контрольная работа должна быть аккуратно оформлена (формат А4, машинописный текст, размер левого поля 20 мм, правого - 10мм, верхнего - 20мм, нижнего 20мм, отступ красной строки 1,5, межстрочный интервал 1,5 шрифт 14, Times New Roman) иметь нумерацию страниц и список использованных источников, в котором указываются все использованные студентом литературные источники, расположенные в алфавитном порядке и пронумерованные.
эссе	Эссе студента - это самостоятельная письменная работа на тему, предложенную преподавателем (тема может быть предложена и студентом, но обязательно должна быть согласована с преподавателем). Цель эссе состоит в развитии навыков самостоятельного творческого мышления и письменного изложения собственных мыслей. Эссе позволяет автору научиться четко и грамотно формулировать мысли, структурировать информацию, использовать основные категории анализа, выделять причинно-следственные связи, иллюстрировать понятия соответствующими примерами, аргументировать свои выводы; овладеть научным стилем речи. Эссе должно содержать: четкое изложение сути поставленной проблемы, включать самостоятельно проведенный анализ этой проблемы с использованием концепций и аналитического инструментария, рассматриваемого в рамках дисциплины, выводы, обобщающие авторскую позицию по поставленной проблеме. В зависимости от специфики дисциплины формы эссе могут значительно дифференцироваться. В некоторых случаях это может быть анализ имеющихся статистических данных по изучаемой проблеме, анализ материалов из средств массовой информации и использованием изучаемых моделей, подробный разбор предложенной задачи с развернутыми мнениями, подбор и детальный анализ примеров, иллюстрирующих проблему и т.д. Структура эссе: - введение (суть и обоснование выбора выбранной темы, краткие определения ключевых терминов); - основная часть (аргументированное раскрытие темы на основе собранного материала); - заключение (обобщения и выводы). Эссе оцениваются по нескольким направлениям: содержание, стиль, способность изложить свои мысли. Основные требования к написанию эссе. – Обозначение круга понятий и теорий, необходимых для ответа на вопрос. – Понимание и правильное использование терминов и понятий. – Использование основных категорий анализа. – Выделение причинно-следственных связей. – Применение аппарата сравнительных характеристик. – Аргументация основных положений эссе. – Наличие промежуточных и конечных выводов. – Личная субъективная оценка по данной проблеме.
экзамен / зачет	При подготовке к экзамену/зачету необходимо опираться, прежде всего, на

	лекции, а также на источники, которые разбирались на семинарах в течение семестра. В каждом билете содержится два вопроса. Ответ предполагает полное и последовательное изложение изученного материала, а также демонстрацию способности и готовности применить полученные теоретические знания к предлагаемым практическим заданиям.

Содержание лабораторных занятий
и темы для самостоятельной работы студентов

Наименование тем (вопросов), изучаемых по даннойдисциплине	**Самостоятельная работа студентов**
Компьютерная лексикография. Первые переводные словари. Виды информации в словаре и в других базах данных.	Работа с онлайн-версиями энциклопедических статей о словарях С. Джонсона, Н. Вебстера и В.И. Даля
Проблемы автоматической обработки текста, необходимой для работы программ, анализирующих и преобразующих текстовые данные. Типология материалов в цифровых массивах.	Анализ словарных статей в онлайн-словарях (представленные в них виды информации) и перекрѐстных ссылок между статьями
Поиск информации как лингвистическая проблема. Групповые проекты. Современные информационно-поисковые системы (Google, Яндекс, Yahoo и др.). Возможности расширенного поиска. Синтаксис запросов. Проблемы машинного перевода. Распределенные вычисления. Перспективы развития компьютерных технологий в филологии.	Анализ сайтов, содержащих статьи и монографии о требованиях к языковому корпусу.
Корпусная лингвистика и требования к корпусу. Специфика разметки языковых данных. Корпуса текстов on-line. Лингвистические принципы автоматического выделения информации из текста	Саморегистрация на сайте национального языкового корпуса и упражнения в отборе информации по определенным параметрам и областям.
Подготовка материалов для учебного процесса. Обучающая среда MOODLE. Ресурсы преподавателей на сайте СОГУ. Курсы по филологии	Особенности онлайновых учебных ресурсов (ознакомление со структурой специализированных сайтов) Анализ методических онлайн ресурсов Интернет, в т.ч. на сайте СОГУ.
Количественные методы в применении к структуре сюжета и стихотворного ритма. Специфика языка художественной литературы. Использование Translation Memory при переводах текстов.	Подготовка материалов для учебного процесса и загрузка их в обучающую среду MOODLE.
Лингвистическая редактура орфографии и грамматики: приемы работы и нерешенные проблемы. Средства конвертирования форматов файлов.	Сопоставительный анализ международных и русскоязычных систем поиска.

Использование ИКТ в научно-исследовательской деятельности..	Анализ проблем и перспектив использования компьютерных технологий в филологии (в т.ч.) машинного перевода.

Теоретический материал, необходимый для изучения темы «Корпусная лингвистика»

Понятийно-категориальный аппарат корпусной лингвистики

Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих теоретических принципов и практических механизмов построения и эксплуатации представительных массивов языковых данных, в том числе корпусов текстов, предназначенных для лингвистических исследований в интересах широкого круга пользователей с применением компьютерных технологий.

В основе корпусной лингвистики лежит то, что язык - это полностью социальное явление, проявляющее себя в текстах, которые можно записать, описать и проанализировать. Внутренние, немые тексты также являются текстами, но их нельзя пронаблюдать и, следовательно, они не являются социальным явлением.

Целесообразность создания и смысл использования корпусов определяется следующими предпосылками: 1) достаточно большой объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений; 2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения; 3) однажды созданный и подготовленный массив данных может использоваться многократно, многими исследователями и в различных целях. Терминология корпусной лингвистики еще не установилась. Во-первых, это естественно, учитывая ее недавнее происхождение.

Во-вторых, корпусная лингвистика как отдельное направление лингвистики сложилась в США и в Великобритании. И соответственно, ее терминология складывалась и продолжает складываться в недрах английского языка. И, естественно, русская корпусная терминология строится на базе англоязычной. В.П. Захаров полагает, что методология корпусной лингвистики может быть применена и к ней самой, то есть, по его мнению, необходимо составить корпус текстов по корпусной лингвистике и разрабатывать словарь по корпусной лингвистике непосредственно на живом текстовом материале.

Некоторое число публикаций на русском языке, посвященных вопросам создания и использования корпусов, уже имеется. Что касается русского языка, то среди специалистов до сих пор нет единодушия в отношении главного термина: «корпус». Каким должно быть множественное число от слова «корпус»? Как образуется соответствующее прилагательное? Словари допускают для разных значений этого существительного две формы множественного числа: «ко́рпусы» и «корпуса́». Для значения «массив», которое имеет место в случае языковых корпусов, именительный падеж множественного числа должен быть «ко́рпусы» и, соответственно, прилагательное «ко́рпусный» (Большой толковый словарь русского языка, СПб., 1998).

Однако анализ узуса специалистов пока свидетельствует в пользу форм «корпуса́», «корпусно́й», «корпусна́я», которые используются заметно чаще, так что можно сказать, что в настоящее время этот вопрос остается открытым. Прежде чем говорить про корпус текстов нужно понять, что такое корпус данных.

Корпус данных представляет собой сформированную по определенным правилам выборку данных из области реализации языковой системы, которая содержит феномены, подлежащие лингвистическому описанию. Область реализации языковой системы, содержащая феномены, подлежащие лингвистическому описанию, называется проблемной об-

ластью. Проблемная область для конкретного корпуса данных может быть сколь угодно велика или мала – все определяется конкретным объектом анализа.

Корпус данных имеет только одно измерение – речевое, поскольку сам по себе он не обладает потенцией производства своих составляющих. Последнее, однако, не означает, что корпус данных не может использоваться для реконструкции языка как системы. Наоборот, это одна из главных задач лингвистического исследования корпуса.

Единица хранения корпуса данных – это некоторая совокупность естественноязыковых выражений проблемной области, для которой составляется одно описание на некотором метаязыке, определяемом процедурой формирования корпуса. Поскольку корпус данных – это некоторая выборка из проблемной области, сформированная по определенным принципам, то единица хранения непосредственно зависит от того по каким принципам осуществлялась выборка. 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 pour Homme”
область косметика и парфюмерия
вид слогана перевод с французского
оригинал Pour les homes qui aiment les femmes qui aiment les hommes
источник «Космополитен»

Выражение «Для мужчин, которые любят женщин, которые любят мужчин» и сопоставленные ему характеристики вместе образуют единицу хранения.

Единицей хранения корпуса по современной публицистике является целый текст, в дескрипции которого зафиксированы следующие характеристики:

источник наиболее полно в корпусе представлены следующие источники: «Век» 8%, «Завтра» 14%, «Известия» 5%, «Итоги» 11%, «Литературная газета» 6%, «Московские новости» 8%, «независимая газета» 6%, «Новый мир» 12%, «Российская газета» 8%, а также другие газеты и журналы;

автор около 1000 авторов;

название статьи 1368 названий;

политическая ориентация издания «общедемократическая» пресса, «левая пресса»

жанр воспоминания, интервью, критика, круглый стол, очерк, проблемная статья, репортаж, рецензия, фельетон;

тема внутренняя политика, внешняя политика, литература, искусство (всего 39 различных тем);

время период 90-х гг., период «ранней перестройки».

Совокупность описаний единиц хранения образует некоторое множество, по которому можно судить о представительности выборки.

Существуют широкое и узкое понимание **корпуса текстов**. В широком понимании корпус текстов – это вид банк данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то отрывки текстов данной проблемной области. Ис ходя из этого определения, любой набор более чем одного текста может быть, в принципе, назван корпусом (от лат. corpus – “body”). Однако понимание корпуса как основы для компьютерной обработки все же отличается от понимания корпуса, как собрания текстов для разного вида литературного и лингвистического анализов, не предполагающих применение специализированного программного обеспечения. В узком смысле под названием лингвистический, или языковой, корпус текстов В.П. Захаров понимает большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. В понятие «кор-

пус текстов» входит также система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего называют корпусным менеджером (или корпус-менеджером) (англ. corpus manager). Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме. В.В. Рыков определяет корпус текстов как некоторое собрание текстов, в основе которых лежит логический замысел, логическая идея, объединяющая эти тексты. Логическая идея воплощается в правилах организации текстов в корпус; алгоритме и программе анализа корпуса текстов; сопряжённой с этим идеологии и методологии.

Порог отображения – соотношение между корпусом данных и проблемной областью при пропорциональном сужении. Корпус данных является сужением проблемной области, при котором некоторые части проблемной области могут оказаться вне корпуса. Например, проблемная область состоит из 20 контекстов, тогда как корпус данных содержит ее четвертую часть – 5 контекстов. Контексты являются примерами реализации различных синтаксических явлений: в 10 контекстах представлены простые предложения, в 8 контекстах – сложные предложения, в 2 последних контекстах – примеры парцелляции. В корпусе данных один контекст соответствует четырем контекстам проблемной области. Это означает, что контексты парцелляции при пропорциональном сужении в четыре раза не попадают в корпус данных. Чем выше порог отображения, т.е. соотношение между корпусом данных и проблемной областью, тем больше вероятность, что какие-то феномены проблемной области не попадут в корпус данных.

Параметризация проблемной области - выделение некоторых характеристик текстов, которые релевантны для предполагаемого исследования. Совокупность этих характеристик (их возможные комбинации) служит основой для отбора текстов в корпус. Например, текстовый корпус Ф. Достоевского создавался как источник для словаря языка писателя, параметры его организации определялись правилами построения словарной статьи: поскольку словарная ста тья предполагала составление указателя ко всем употреблениям слова, то корпус должен был охватывать все тексты Ф. Достоевского.

В отличие от корпуса текстов Ф. Достоевского, корпус по современной публицистике потребовал разработки сетки параметров, позволяющих осуществить инвентаризацию проблемной области и обеспечить ее репрезентативное представление в корпусе. Параметризация проблемной области при формировании корпуса текстов по современной публицистике основывается на следующих основных факторах:

− фактор автора текста: журналист / непрофессиональный политик или профессиональный политик (распределение по политикам учитывает как крупных политических деятелей типа Ельцина, Путина, Черномырдина, Немцова, Хакамады, Селезнева, Гайдара, Жириновского, так и политиков второго ряда); отдельно стоит проблема выявления «команд спичрайтеров», определяющих собственно языковое оформление текста;

− фактор персонификации-деперсонификации автора (конкретный человек или партия / общественное движение / политическая организация / учреждение или деперсонифицированный текст – лозунги, передовицы и т.п.);

− фактор адресата (сторонники – противники – нейтральная аудитория; профессиональная ориентация – выступление перед шахтерами, творческой интеллигенцией и пр.);

− фактор прагматических условий порождения текста (речь на митинге – речь на заседании институционального органа – интервью – пресс-конференция, всего было учтено 15 типов условий произнесения);

− фактор источника (журнальный текст – книжный текст – листовка – агитационный плакат – лозунг – телевидение – радио);

− фактор коммуникативного распределения (монологический текст – диалог; общие типы иллокуций: демонстрация намерений, например, политическая программа – аргументативный диалог и пр.).

На основе перечисленных факторов были сформированы параметры, позволившие выделить из проблемной области около 70 типов текстов. Созданный корпус текстов по современной русской публицистике с точки зрения выбранных параметров может рассматриваться как модель функционирования языка современной публицистике в дискурсе. Требования к корпусу текстов, предъявляемые пользователями, включают следующие параметры: репрезентативность, полнота, экономичность, структуризация материала, самодостаточность, компьютерная поддержка.

Под **репрезентативностью** понимается способность корпуса текстов отражать все свойства проблемной области, релевантные для данного типа лингвистических исследований, в установленной пропорции, определяемой частотой явления в проблемной области. Кроме того, репрезентативность включает необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. Можно сказать, что применительно к общеязыковому (национальному) корпусу это понятие невозможно рассчитать и описать строго математически, однако к этому можно и нужно стремиться, как на этапе проектирования корпуса, так и на этапе его эксплуатации. Задача создателей корпуса – собрать как можно большее количество текстов, относящихся к той предметной области, для изучения которой корпус создается. Но главное не только и не столько в количестве языкового материала, сколько в его пропорциональности. Можно сказать, что корпус – это уменьшенная модель языка или подъязыка.

Если репрезентативность указывает на то, что единицы проблемной области отражаются пропорционально в корпусе данных, то полнота требует учета релевантных явлений, даже если это не соответствует идее пропорционального сужения, поскольку при определенном пороге некоторые явления могут исчезнуть из корпуса.

Экономичность корпуса предполагает, что корпус текстов является не просто строгим подмножеством текстов проблемной области, но, по возможности, существенно отличаться от нее по объему. Однако, чем более экономичен корпус, тем выше порог отображения, тем более вероятно, что в корпус данных не попадут малочастотные явления, поэтому экономия не может проводиться в ущерб репрезентативности.

Структуризация материала заключается в обеспечении корпуса описью данных, в которых единицы хранения характеризуются по тем параметрам, которые могут оказаться важными для пользователя.

Самодостаточность – свойство не корпуса в целом, а его единиц хранения, на которые могут быть наложены существенные ограничения, т.е. единицей хранения может оказаться не целый текст, а его фрагмент (предложение или группа связанных между собой предложений), который не должен содержать неоднозначности любых типов, например, местоимений, для которых невозможно восстановить антецедент и пр. В случаях, когда единицы хранения включают случаи языковой игры, связанной с неоднозначностью, рамки контекста должны быть таковы, чтобы пользователь мог легко ее распознать.

Что касается **компьютерной поддержки**, то желательно, чтобы корпус текстов имел комплексное программное обеспечение по обработке данных, обеспечивающих ряд функций, а именно:

− стандартный просмотр текстов, входящих в состав корпуса (в виде просмотра таблицы базы данных);

− выборка и упорядочивание текстов по указанным формальным или содержательным признакам, а также на основе их различных комбинаций;

− получение текстовых массивов, являющихся подмножествами исходного корпуса на основе изменяемой случайной выборки и задания процентов от общего корпуса по одному из параметров;

− просмотр массивов-подмножеств и их обработка;

− поиск конкретных словоформ;

− поиск словоформ по леммам – словарным формам (лемматизация – приведение словоформ к словарной форме);

− поиск группы словоформ в виде разрывной или неразрывной синтагмы;

− поиск словоформ по набору морфологических признаков;

− отображение информации о происхождении, типе текста и т.п.;

− вывод результатов поиска с указанием контекста заданной длины;

− статистическая инвентаризация (получение различных статистических данных о языковых и речевых явлениях, например, данные о частоте словоформ, лексем, грамматических категорий, данные о совместной встречаемости лексических единиц и т.д.);

− составление конкордансов (список всех употреблений данного слова в контексте со ссылками на источник);

− автоматическая словарная обработка (составление полных и частичных словников по различным критериям – частоте, алфавиту и пр.);

− сохранение отобранных строк конкорданса в отдельном файле на компьютере пользователя и др.

Принципы классификации корпусов

Существует большое число разных типов корпусов. Их разнообразие определяется многообразием исследовательских и прикладных задач, для решения которых они создаются, и различными основаниями для классификации. В зависимости от поставленных целей и классифицирующих признаков, можно выделить различные типы корпусов:

1) **по форме хранения** (по типу данных): в звуковой форме (речевые), письменные, смешанные;

2) **по параллельности**: одноязычные; двуязычные; многоязычные. Каждый из выделенных по принципу параллельности типов, в свою очередь, классифицируется по языку (языкам) представления текстов;

3) **по так называемой «литературности»**: литературные, диалектные, разговорные, терминологические, смешанные;

4) **по жанровой принадлежности**: художественные, фольклорные, драматургические, публицистические, смешанные;

5) **по «общности»**: разных авторов, одного автора.

Классификации по таким критериям, как литературность, жанровая принадлежность, в сущности, отражают единый принцип противопоставления общности корпусов, относящихся ко всему языку, корпусам, относящимся к какому-либо подъязыку (жанр, стиль, язык определенной возрастной или социальной группы, язык писателя или ученого и т.п.).

6) **по способам доступа**: свободно доступные, коммерческие, закрытые;

7) **по назначению**: исследовательские, иллюстративные, параллельные (выделенные в данной группе В.В. Рыковым, тогда как другие исследователи называют их двух- и многоязычными и относят к типам корпусов, дифференцированным по принципу параллельности). Исследовательскими называются такие корпусы, которые предназначены преимущественно для изучения раз личных аспектов функционирования языковой системы. Они строятся не после проведения какого-либо исследования, а до него. Иллюстративные корпусы создаются после проведения научного исследования: их цель не столько выявить новые факты, сколько подтвердить и обосновать уже полученные результаты (корпус дискурсивных слов русского языка для обеспечения примерами одноименного словаря). Корпусы параллельных текстов по структуре представляют собой подмножество текстов на языке-источнике и одно или несколько подмножеств текстов, которые являются переводами текстов языка-источника на языки-цели. Подобные корпусы формируются для научных и практических целей, в частности, для преподавания иностранных языков. Например, английский текст “Alice in Wonderland” и его переводы на немецкий, французский и русский языки могут формировать такой корпус или быть частью большего корпуса параллельных текстов;

8) **по способу существования**: динамические (неограниченные, мониторные), статические (ограниченные). Статические корпусы не развиваются и отражают определенное состояние языковой системы. Однако значительная часть чисто лингвистических и не только лингвистических исследований требует выявления функционирования языковых феноменов на временной шкале – например, изменения значения слов, частоты использования тех или иных синтаксических конструкций. Для отражения процессуального аспекта проблемной области была разработана новая технология построения и эксплуатации динамического корпуса текстов, который еще называется мониторным. Динамические корпусы не предполагают раз и навсегда заданного набора текстов. В течение заранее фиксированного промежутка времени происходит обновление и/или дополнение множества текстов корпуса;

9) некоторые исследователи разграничивают такие критерии, как способ существования самого корпуса и способ хронологической репрезентации данных о языке. **По хронологическому аспекту** выделяются следующие разновидности корпуса: синхронический (фиксирует определенное состояние языка); мониторный (отслеживает текущее состояние языка); диахронический (представляет произошедшие в языке изменения);

10) **по объему текстов**: полнотекстовые, «фрагментнотекстовые»;

11) **по наличию дополнительной информации** (по индексации): аннотированные (размеченные), простые (неразмеченные). Указанные два типа корпусов отличаются по способам представления данных, основанным на современных компьютерных технологиях. Если простые корпусы основаны на неструктурированном формате хранения (запись графем текста в ASCI кодах), то аннотированные корпусы основаны на структурированном формате хранения (текст со специальной разметкой). Существует мнение, что тексты представленные в неструктурированном формате хранения, т.е. не содержащие разметки, корпусом не считаются, поскольку разметка - главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов, в изобилии представленных в современном интернете, которые несмотря на академический режим подачи текстов, максимально точное воспроизведение авторитетных печатных изданий, в необработанном виде для научных исследований языка пригодны очень ограниченно. Кроме того, библиотеки создаются теми, кому интересно в большей степени содержание текстов, чем их языковые качества. Для составителей корпуса такие факторы, как увлекательность или полезность книги, ее высокие художественные или научные достоинства являются важными, но не первостепенными. Корпус, в отличие от электронной библиотеки, — это не собрание «интересных» или «полезных» текстов; это собрание текстов, интересных или полезных для изучения языка. А такими могут оказаться и роман второстепенного писателя, и запись обычного телефонного разговора, и типовой договор аренды и т.п. — наряду, конечно, с классическими произведениями художественной литературы. Иными словами, неаннотированных корпусов не существует.

12) **по характеру разметки**: морфологические, синтаксические, семантические, просодические и т.д.

Аннотирование корпусов. Виды корпусной разметки

Для решения различных лингвистических задач мало лишь наличия массива текстов. Требуется также, чтобы тексты содержали в себе явным образом разного рода дополнительную лингвистическую и экстралингвистическую информацию. Так в корпусной лингвистике возникла идея аннотирования (разметки) корпуса и появились аннотированные (размеченные) корпусы. Действительно, уже на уровне статистических подсчетов можно получить более интересные результаты, если вместе с каждым словом хранится информация о его частеречной принадлежности: появляется возможность подсчитывать не просто частотность слов, а частотность представителей тех или иных частей речи.

Разметка (аннотирование, тэггинг) (от англ. tagging, annotation) заключается в приписывании текстам и их компонентам специальных кодов, меток, тэгов (от англ. tag – ярлык,

метка), каждому из которых соответствует определенный набор признаков, характеризующих данный текст или его фрагмент. Разметка, которую может содержать корпус, подразделяется на лингвистическую и экстралингвистическую, последняя включает так называемую метаразметку и структурную разметку.

Собственно лингвистическая разметка описывает лексические, грамматические и прочие характеристики элементов текста.

Метаразметка касается сведений об авторе и тексте. Причем сведения об авторе могут включать не только его имя, но также и возраст, пол, годы жизни и многое другое, а сведения о тексте обычно содержат, кроме названия, еще и язык, на котором он написан, год и место издания, жанр, тематику и т.д. Наличие подобной информации позволяет значительно детализировать поиск в текстовых базах данных и, кроме того, предоставляет средства идентификации соответствующего документа.

Структурная разметка отражает особенности форматирования текста (заголовки, главы, абзацы, отступы, предложения, слофоформы и т.д.).

Набор этих данных во многом определяет возможности, предоставляемые корпусами исследователям. При их выборе необходимо руководствоваться целями исследования и потребностями лингвистов, а также возможностями по внесению в текст тех или иных дополнительных признаков.

Среди лингвистических типов разметки выделяются:

− **морфологическая разметка**. В иностранной терминологии употребляется термин part-of-speech tagging (POS-tagging), дословно – частеречная разметка. В действительности морфологические метки включают не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. Это основной тип разметки: во-первых, большинство крупных корпусов являются как раз морфологически размеченными, вовторых, морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического, и, в-третьих, успехи в компьютерной морфологии позволяют автоматически размечать корпусы больших размеров. Схема морфологической разметки предполагает наличие, во-первых, набора тэгов, во-вторых, описания того, что каждый из них означает и, в-третьих, правил присвоения тэгов единицам текста. Размер наборов тэгов, применяемых в разных корпусах, варьируется. Несомненно, чем больше набор тэгов, тем более детальный анализ текста осуществим с его помощью;

− **синтаксическая разметка**, являющаяся результатом синтаксического анализа, или парсинга (англ. parsing), выполняемого на основе грамматики структур непосредственно составляющих. Графически синтагматические отношения между членами предложения изображаются, как известно, в виде дерева, а в тексте они представлены парами из открывающейся и закрывающейся квадратных скобок, которые обрамляют различные синтаксические конструкции – именные, глагольные и предложные словосочетания, придаточные предложения. Рядом как с открывающейся, так и с закрывающейся скобкой ставятся метки (коды), описывающие заключенную в них конструкцию. Одни пары скобок вложены в другие, элементом высшего уровня является предложение. Этот вид разметки описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции. Тексты, получившие синтаксическую разметку, известны как treebanks.

− **семантическая разметка**. Хотя для семантической, как и для других видов разметки, нет стандартной формы, чаще всего для ее представления используют код, состоящий из букв и цифр или только цифр, в котором первая буква или цифра обозначает общую семантическую категорию, в которую входит данное слово, а последующие символы – более узкие подкатегории, специализирующие его значение. В схемах семантической разметки предусмотрены те случаи, когда в качестве единицы смысла выступает не отдельное слово, а словосочетание. Все члены такого словосочетания получают один и тот же код, при этом для каждого из них дополнительно указываются его порядковый номер, а также общее число слов в идиоматическом выражении; − анафорическая разметка.

Из всех видов референции наибольшую сложность для автоматической обработки текста представляет местоименная. Так, большинство систем машинного перевода обрабатывает текст по отдельным предложениям, отчего страдает связность выходного текста. Эффективность таких систем гораздо повысилась бы, если бы правильно определялась референция местоимений-заместителей. Этому и призвана способствовать анафорическая разметка, которая фиксирует референтные связи, в частности, местоименные. Антецедент, в роли которого обычно выступает именное словосочетание, берется в пронумерованные скобки, а рядом с местоимением-заместителем ставится особый знак, отсылающий к антецеденту с соответствующим номером; − просодическая разметка. В просодических корпусах применяются метки, описывающие ударение и интонацию. В корпусах устной разговорной речи просодическая разметка часто сопровождается так называемой дискурсной разметкой, которая служит для обозначения пауз, повторов, оговорок, и т.д.

Существуют и другие типы разметки.

На синтаксическом уровне, как и на морфологическом, проявляется тенденция к меньшей детализации схем грамматической разметки в целях увеличения скорости и последовательности анализа текста. Метод синтаксической разметки, который возник в результате этой тенденции, получил название skeleton parsing.

В.П. Захаров считает, что несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа. При этом следует подчеркнуть, что корпус с синтаксической разметкой явно или неявно включает в себя и морфологические характеристики лексических единиц.

Разметка корпусов представляет собой трудоемкую операцию, особенно учитывая размеры современных корпусов. Она может осуществляться автоматически и вручную. Фактически, корпус в его современном понимании – это всегда компьютерная база данных, и в процессе его создания и эксплуатации естественно использование специальных программ. Среди этих программ особое место занимают программы автоматической разметки. Их использование экономичнее с точки зрения временных и трудозатрат. Для морфологической и синтаксической разметки существуют различные программные средства, которые принято называть соответственно тэггеры (taggers) и парсеры (parsers). В результате работы программ автоматического морфологического анализа каждой лексической единице приписываются грамматические характеристики, включая часть речи, лемму (нормальную форму) и набор граммем (например, род, число, падеж, одушевленность/неодушевленность, переходность и т.п.). В результате работы программ автоматического синтаксического анализа фиксируются синтаксические связи между словами и словосочетаниями, а синтакси ческим единицам приписываются соответствующие характеристики (тип предложения, синтаксическая функция словосочетания и т.п.). Большинство таких систем все же требует ручного постредактирования, так как в случаях морфологической омонимии и синтаксической неоднозначности программа предлагает несколько вариантов решения, из которых нужный выбирает исследователь. Однако корпусы нового поколения включают сотни миллионов слов, поэтому выдвигаются принципы разработки систем, которые бы минимизировали вмешательство человека путем автоматического разрешения морфологической или синтаксической омонимии.

Корпусы, как правило, предназначены для многократного использования многими пользователями, соответственно, и их разметка, и их программное обеспечение должны быть определенным образом унифицированы. Что касается разметки, то как лингвистическая, так и экстралингвистическая разметка должны базироваться на некоторых достаточно широко распространенных и принятых принципах описания текстов и языковых единиц. Параметры разметки и их значения должны быть достаточно «естественными», т.е. должны соответствовать общепринятым научным классификациям. Что касается программного обеспечения, то оно должно поддерживать обработку типовых запросов и решение типовых задач. Большое значение имеет унификация форматов представления дан-

ных, как их наполнения, так и структуры. Единые форматы представления данных позволяют во многих случаях использовать единое программное обеспечение и обмениваться корпусными данными. Стандартизация в отношении корпусов, совместимость типов данных важны и с точки зрения сравнимости разных корпусов. Вопросы оценки корпусов, их пригодности к различным заданиям также требуют своих «стандартов оценки».

В настоящее время не существует общепризнанных стандартов представления информации в текстах. Специальный международный проект Text Encoding Initiative (TEI) предназначен для того, чтобы разработать стандартизированные средства разметки руководствуясь рекомендациями EAGLES (Expert Advisory Group on Language Engineering Standards). В качестве формального языка разметки широко применяется язык SGML и его подмножество XML, которые получили статус общепризнанных международных языков разметки документов. В настоящее время стандарты EAGLES непосредственно включаются в технологическую среду языка XML, примером чего может служить, в частности, разработка стандарта Corpus Encoding Standard for XML (XCES).

Конструирование и применение корпусов. Лингвистические исследования на базе корпуса

При конструировании и применении корпусов единой методики для всех языков нет, так как различаются языки, традиции, технологические процессы.

Технологический процесс создания корпуса можно представить в виде следующих шагов или этапов.

1. Определение пользователей корпуса, логической идеи, положенной в его основу, перечня источников, объёма данных его необходимости и реалистичности, единиц хранения корпуса (отрывки текста, полные тексты или и то и другое).

2. Оцифровка текстов (преобразование в компьютерную форму). Следует сказать, что насколько раньше задача ввода текстов в компьютер была тяжела и трудоемка, настолько сегодня эта проблема решается довольно легко, по крайней мере, что касается современных текстов и современной орфографии. Эта легкость базируется на успехах в оптическом вводе (сканирование) и распознавании текстовой информации и на глобальной компьютеризации современной жизни, в том числе и в областях, связанных с обработкой текстовой информации. Тексты в электронном виде для создания корпусов могут быть получены самыми разными способами - ручной ввод, сканирование, авторские копии, Интернет, оригинал-макеты, предоставляемые составителям корпусов издательствами и пр.

3. Предобработка текста. На этом этапе все тексты, полученные из разных источников, проходят филологическую выверку и корректировку. Также осуществляется подготовка библиографического и экстралингвистического описания текста.

4. Конвертирование и графематический анализ. Некоторые тексты проходят также через один или несколько этапов предварительной машинной обработки, в ходе которых осуществляются различного рода перекодировка (если требуется), удаление или преобразование нетекстовых элементов (рисунки, таблицы), удаление из текста переносов, разрывов строк, обеспечение единообразного написания тире и пр. Как правило, эти операции выполняются в автоматическом режиме. Обычно на этом же этапе осуществляется сегментирование текста на его структурные составляющие.

5. Разметка текста. Структурная разметка текстов (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически. Метаразметка, включающая помимо содержательных данных (библиографические сведения, признаки, характеризующие жанровые и стилевые особенности текста, информацию об авторе), еще и дополнительные формальные данные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ), обычно вводится вручную.

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

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

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

Конечно, в каждом конкретном случае состав и количество процедур могут отличаться от перечисленных выше, и реальная технология может оказаться гораздо сложнее.

Пользователей корпусов, как правило, интересует не содержание конкретных текстов, а их метатекстовая информация и примеры употребления тех или иных языковых элементов и конструкций. Это, в первую очередь, лингвисты. Первоначальные лингвистические исследования, проводившиеся с помощью корпусов, сводились к подсчету частот встречаемости различных языковых элементов. Чаще всего этими элементами были слова, в других случаях – графемы, морфемы, словосочетания. Статистические методики используются в решении сложных лингвистических задач в области машинного перевода, автоматической проверки орфографии и грамматики и т.д. Так, устойчивые словосочетания представляют собой с семантической точки зрения неделимую смысловую единицу, что очень важно учитывать в лексикографии, системах автоматической обработки текста. На материале корпуса статистическими методами можно определить, какие слова встречаются вместе регулярно и, таким образом, могут быть отнесены к устойчивым словосочетаниям.

По-прошествии времени корпусы стали осознаваться как мощные информационные ресурсы, могущие быть использованными в рамках различных лингвистических направлений. Так, корпусы являются богатым источником данных для многоаспектых лексикографических работ по подготовке разнообразных исторических и современных словарей. На основе корпуса и с применением компьютера словари могут составляться и пересматриваться гораздо быстрее, чем раньше, таким образом, фиксируя текущее состояние языка и не успевая устаревать за то время, которое проходит от момента начала работы над ними до момента выхода их из печати. Представительный массив языковых данных за определенный период позволяет отслеживать неологизмы, изменение значений у уже известных слов, варьирование частот и контекстов в различные периоды времени. С исследованиями по лексикографии тесно связаны исследования в области семантики. Наблюдая окружения той или иной лингвистической единицы в корпусе, можно установить определенные семантические признаки, характеризующие данную единицу. Часто слово входит сразу в несколько семантических категорий, поэтому степень его принадлежности к той или иной категории может быть выявлена путем подсчета частот его распределения по разным категориям. Данные корпусов могут быть использованы для построения и уточнения грамматик, проведения анализа лексикограмматических характеристик в разных жанрах, у разных авторов и т.д