

Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра инженерной физики

О.П. Оганджян

**МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ
ПО ИЗУЧЕНИЮ ДИСЦИПЛИНЫ
«ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
В ФИЛОЛОГИИ»**

*Витебск
ВГУ имени П.М. Машерова
2014*

УДК 004.4(075.8)
ББК 32.973я73
О-36

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 6 от 25.06.2014 г.

Автор: старший преподаватель кафедры инженерной физики ВГУ имени П.М. Машерова **О.П. Оганджян**

Рецензент:
заведующий кафедрой иностранных языков
ВГУ имени П.М. Машерова, кандидат филологических наук,
доцент *Д.О. Половцев*

Оганджян, О.П.
О-36 Методические рекомендации по изучению дисциплины «Информационные технологии в филологии» / О.П. Оганджян. – Витебск : ВГУ имени П.М. Машерова, 2014. – 50 с.

В методических рекомендациях по дисциплине «ИТ в филологии» рассмотрены технологии семантического поиска, работы с лингвистическими информационными ресурсами и создания электронных словарей. Призваны оказать помощь студентам, обучающимся по модульной технологии, в освоении учебного материала, его систематизации и подготовке к лабораторным и практическим занятиям, итоговому контролю знаний.

Данное издание предназначено для студентов, преподавателей данной дисциплины и всех интересующихся вопросами информационных технологий в филологии.

УДК 004.4(075.8)
ББК 32.973я73

© Оганджян О.П., 2014
© ВГУ имени П.М. Машерова, 2014

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	4
1. РЕКОМЕНДАЦИИ ПО ПРАКТИЧЕСКОЙ ЧАСТИ	5
2. РЕКОМЕНДАЦИИ ПО САМОСТОЯТЕЛЬНОЙ РАБОТЕ	47
3. ТЕМАТИКА РЕФЕРАТОВ	47
4. ПРИМЕРНЫЕ ВОПРОСЫ К ЗАЧЕТУ	48
ЛИТЕРАТУРА	49

ВВЕДЕНИЕ

Методические рекомендации составлены в соответствии с учебной программой дисциплины «Информационные технологии в филологии». Данные материалы предназначены для проведения лабораторных и практических занятий со студентами, изучающими данную дисциплину, а также для самостоятельной работы над ее содержанием. Целью данного издания является закрепление и углубление теоретических знаний в области информационных технологий, а также формирование и развитие у студентов навыков их практического применения в профессиональной деятельности.

Методологической особенностью дисциплины «ИТ в филологии» является принцип активного творческого мышления, направленность обучения на формирование самостоятельности суждений. Поэтому методические рекомендации содержат задания, конкретные примеры, проблемные вопросы. Использование данных материалов в учебной деятельности позволит активизировать процесс усвоения студентами теоретических положений дисциплины и приобрести необходимые практические навыки. Предложенные практические задания построены на использовании лингвистических информационных ресурсов, электронных словарей, терминологических баз данных. Для анализа предложенных в учебном издании конкретных примеров необходимы знания как теории компьютерной лексикографии, терминографии и баз данных в целом, так и отдельных программных продуктов, необходимых для решения конкретных задач. Тесты – это специальный тип заданий, содержащий несколько вариантов ответа на поставленные вопросы. Для самостоятельной подготовки студентов в соответствии с учебно-тематическим планом вынесено четыре темы. В данном учебном издании предложены задания для них по этим темам. Творческие задания рассчитаны на совместный поиск идей, поэтому их лучше всего выполнять в микрогруппах с последующей защитой подготовленных проектов. Использование данной методики способствует созданию творческой обстановки на занятиях и развитию навыков интеллектуального сотрудничества.

Настоящие рекомендации призваны оказать помощь студентам в закреплении учебного материала и его систематизации.

1. РЕКОМЕНДАЦИИ ПО ПРАКТИЧЕСКОЙ ЧАСТИ

Модуль. Базы данных и лингвистические информационные ресурсы (ЛИР).

Учебные элементы модуля:

УЭ-1. Internet как коммуникационный, информационный, учебный и научно-исследовательский ресурс. Лингвистические информационные ресурсы. УЭ-2. Работа над проектом «Создание БД «Словарь морфем английского языка»».

Тема работы: Информационный поиск в Интернет. Компьютерная лексикография и терминография. Корпусная лингвистика и машинный перевод.

Цель работы: Освоить технологию автоматического анализа текста, синтаксического (простого, с подстановочными знаками) и семантического поиска лингвистической информации в ИПС Google, AskNet и с помощью корпусных технологий. Работа с электронными словарями и терминологическими банками данных.

Ход работы

1. Скачать [описание работы](#).
2. Выполнить задания, предложенные в работе.
3. Результаты выполнения работы (Фамилия.docx) отправить преподавателю на проверку.
4. Ответить на вопросы преподавателя (устно на занятии).

Описание работы

Перед выполнением заданий прочитайте основные теоретические сведения.

Теоретические сведения

Mozilla Firefox, Opera, Google Chrome – программы-браузеры для навигации в WWW и просмотра Web-страниц.

Web-страница – документ, содержащий форматированный текст, мультимедийные объекты (графика, звук, видео), **ссылки** на другие Web-страницы или иные ресурсы Internet. **Гипертекстовая ссылка** – выделенная часть документа, реализующая переход к другому документу.

Адресная строка браузера используется для ввода адреса (URL) ресурса, к которому необходимо получить доступ. Если при этом содержимое загружаемого ресурса не помещается в отведенное поле просмотра, то в центральной части рабочего окна программы появляются вертикальная и горизонтальная полосы прокрутки.

URL (универсальный указатель ресурсов) – адрес любого файла в Internet. В URL содержится название протокола, по которому нужно обращаться к файлу, адрес компьютера с указанием, какую программу-сервер запустить на нем, и полный путь к файлу.

Пример URL:

http://www.fio.by/predmet/inostr_vaz.php – адрес полезных ресурсов по иностранным языкам, расположенного в папке predmet на Web-сервере Республиканского центра Интернет-образования.

Синтаксический и семантический поиск в ресурсах Интернет

Традиционными способами поиска информации человеком являются:

- поиск «сверху» (по оглавлению);
- Поиск «снизу» (с помощью различных указателей);
- поиск с помощью гипертекстовых связей (перекрестных ссылок);
- полнотекстовый поиск путем просмотра всего текста.

Последний вид поиска является наиболее точным, но и наиболее трудоемким, требующим больше всего времени и усилий. Организация поиска предполагает следующие составляющие и этапы:

- 1) множество документов (текстов или их фрагментов), по которым следует производить поиск;
- 2) коммуникативная потребность в информации, выражающаяся в информационном запросе пользователя;
- 3) удовлетворение коммуникативной потребности, состоящее в выборе той части текстов исходного массива, которая соответствует информационному запросу.

Упорядоченная совокупность документов и информационных технологий, предназначенных для хранения и поиска информации, представленной в виде текстов или их частей (фактов), получила название *информационно-поисковой системы (ИПС)*.

К наиболее известным ИПС относятся:

- Excite (www.excite.com);
- Yahoo! (www.yahoo.com);
- MSN (www.msn.com);
- Google (www.google.ru);
- Яндекс (www.yandex.ru).

Пользователь вводит свой поисковый запрос в специальную строку ИПС. Этот запрос, сформулированный на естественном языке, программой поиска преобразуется в *информационно-поисковый язык (ИПЯ)* – формальный язык, предназначенный для описания содержания документов, хранящихся в ИПС, и запроса. Процедура описания документа на ИПЯ называется индексированием. В результате индексирования каждому документу приписывается его формальное описание – поисковый образ документа. Аналогичным образом индексируется и запрос, которому приписывается поисковый образ запроса или поисковое предписание. Алгоритмы информационного по-

иска основаны на сравнении поискового предписания с поисковым образом запроса.

Степень соответствия документа запросу задается категорией **релевантности**¹. При этом в процессе информационного поиска можно получить в выдаче значительный информационный шум – множество документов, формально релевантных, но не являющихся релевантными по смыслу.

Результаты поиска могут характеризоваться с двух точек зрения полноты и точности. *Полнотой поиска* (англ. Recall) называется мера, вычисляемая как отношение количества выданных релевантных документов к общему числу релевантных документов, содержащихся в информационном массиве. *Точность поиска* (англ. Precision) – это отношение количества выданных релевантных документов к общему числу документов в выдаче.

Составить представление о полноте и точности поиска можно, сравнивая выдачи разных поисковых систем. При четком определении ключевых слов запроса и их синтаксической связи значения полноты и точности поиска будут стремиться к единице, т.е. к минимуму релевантных документов, что облегчает выбор человеком нужного результата поиска.

Семантический поиск – поиск информации по смыслу, наиболее соответствующей введенной фразе. Человек, после прочтения содержимого текста, может отнести его к определенной глобальной группе тем: об автомобилях, природе, технологиях, медицине и к более конкретной локальной группе: ремонт трансмиссии автомобиля, поведение насекомых, филология. В зависимости от интеллекта, уровня эрудиции, конкретного человека может быть определена и субъективная мера близости по смыслу текстовых документов.

Синтаксический поиск является первым и до сих пор основным видом информационного поиска. Он основан на чисто синтаксическом соответствии количества слов (лексем) в запросе и идентичных слов в искомых документах.

Современные поисковые системы учитывают также локальную частоту лексем в документе, расположение лексем, частоту встречаемости поискового набора слов во всех соответствующих документах, и многое другое.

Современное состояние проектов семантических поисковых систем и технологий (<http://semanticus.ru/> и <http://asknet.ru/>) характеризуется наличием нескольких подходов:

- использование диалогового пользовательского интерфейса,

¹ При поиске в Интернет важны две составляющие – полнота (ничего не потеряно) и точность (не найдено ничего лишнего). Обычно это все называют одним словом – релевантность, то есть соответствие ответа вопросу.

- категориальные технологии,
- внедрение семантической разметки веб-страниц.

Семантический поиск в Семантикус (<http://semanticus.ru/>) основан на вычислении семантического расстояния между поисковой фразой и материалом, найденным поисковой машиной, сортировкой, ранжированием, количественным и вербальным описанием материалов поиска.

Семантический поиск является несомненно более дорогим (по затратам ресурсов вычислительной системы и времени решения задачи), но более точным инструментом информационного поиска.

Задание 1.

1. С помощью любой информационно-поисковой системы (Rambler, Yandex, Google и др.) найдите в Интернете текст Alice's Adventures in Wonderland by Lewis Carroll tell (например, на сайте www.gutenberg.org/ebooks/11). Сохраните его на свой компьютер в формате MS Word.

2. Выполните поиск с подстановочными знаками по тексту Alice's Adventures in Wonderland by Lewis Carroll tell. Внесите результаты поиска в таблицу. Сохраните таблицу 1 в файл Poisk.doc.

Таблица 1

Задание	Формула поиска	Ответ
1. Найдите в тексте первые пять слов, состоящих из пяти букв		
2. Сколько в тексте шестибуквенных слов, начинающихся на букву s и заканчивающиеся на букву r?		
3. Найдите в тексте первые пять трёхбуквенных слов, начинающиеся на гласную букву		
4. Сколько в тексте слов, состоящих из двенадцати букв? По каким формальным признакам их можно сгруппировать? Приведите пример из каждой группы слов		Группы:

5. Сколько в тексте слов с суффиксом -tion? Приведите пример использования такого слова в контексте		Пример:
6. Есть ли в тексте слова, включающие четыре согласные буквы подряд?		
7. Сколько раз в тексте встречаются пассивные конструкции единственного числа прошедшего времени?		

3. Ознакомьтесь с информационно-поисковым языком двух поисковых систем: Google и Рамблер, которые вы можете найти по ссылкам www.google.ru/intl/ru/help/refinerearoh.html и <http://help.rambler.ru/project.html?s=search>

4. Используя сведения об особенностях ИПЯ каждой поисковой системы, сформулируйте запрос, по которому вы сможете найти информацию, где и когда появился термин «лингвистика». Сравните информационно-поисковые системы по качеству поиска.

Таблица 2

Параметр	Google	Рамблер
Запрос		
Документ, отвечающий результатам запроса (url)		
Номер этого документа в списке результатов		
Инф. шум (количество нерелевантных ссылок)		
Полнота		
Точность		
Выводы (результаты какой ИПС были более полными и точными, где было меньше информационного шума, синтаксис какой ИПС более комплексный, простой, удобный):		

5. Откройте страницу семантической поисковой системы <http://asknet.ru/>

6. Ознакомьтесь с технологией поиска.
7. Сравните результаты поиска в Яндекс.ру и в Аскнет.ру по запросу «Тема детства в произведениях Марка Твена».
8. Сравните эти поисковые системы по релевантности.

Таблица 3

Релевантность	Яндекс.ру	Аскнет.ру
Полнота		
Точность		

Поиск лингвистической информации с помощью корпусных технологий.

Для филологических исследований одной из важных задач является сбор и хранение источников фактического материала. В настоящий момент для решения этой задачи используются большие собрания текстов самой разной функциональной направленности, которые удобно хранить в электронном виде. Привлечение компьютеров и специальных телекоммуникационных сетей позволяет не только сохранять большие объемы текстов в электронном виде, но и осуществлять поиск по ним, обрабатывать их и т.п. Задача создания корпусов текста является значимой для современной лингвистики, а сами корпуса текстов становятся объектом исследований корпусной лингвистики.

Корпусная лингвистика – раздел прикладной лингвистики занимающийся разработкой общих принципов построения и использования лингвистических корпусов при помощи компьютеров.

Исходя из этого, корпусная лингвистика включает два аспекта:

1. создание корпусов текстов с автоматическими инструментами их использования;
2. разработка способов экспериментальных исследований различных уровней языка на базе корпусов разных типов.

Кроме проведения научных исследований корпусы могут использоваться:

- 1) в лексикографии для создания словарей, определения значения многозначных слов и т.д.;
- 2) в грамматике для определения частоты морфем, типов словосочетаний и предложений и т.д.;
- 3) в лингвистике текста для дифференциации типов текста, выявления связей внутри абзаца и между абзацами и т.д.;
- 4) в автоматическом переводе текстов для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов в параллельных текстах и т.д.;

5) в учебных целях для выбора цитат, фрагментов произведений, примеров для организации учебных занятий, создания учебных пособий и т.д.;

б) в тестировании программ автоматического анализа и синтеза речи и т.д.

Центральное понятие корпусной лингвистики — лингвистический корпус — определяется как совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска. Таким образом, корпус можно кратко охарактеризовать следующим образом:

Корпус = тексты + их разметка.

Практически все корпуса являются лингвистически размеченными.

В качестве подобных неразмеченных корпусов можно рассматривать существующие электронные коллекции текстов: виртуальные библиотеки, архивы электронных версий периодических изданий или новостных лент, которые оказываются достаточными для некоторых исследовательских и учебных целей. Но использование неразмеченных собраний текстов, имеющих инструменты поиска, повышает долю информации, которая может оказаться нерелевантной для исследователя, что значительно затрудняет работу с таким источником. В связи с этим предметом корпусной лингвистики являются преимущественно размеченные корпуса текстов.

Первым этапом в создании корпуса является отбор текстов. При этом следует продумать, тексты каких функциональных стилей и конкретных жанров, какого года издания и в каком количестве будут включены в корпус. При отборе текстов в корпус следует ориентироваться на следующие требования к созданию корпусов:

1) *репрезентативность* (частота явления в корпусе должна соответствовать его частоте в естественном языке);

2) *полнота* (явление должно включаться в корпус, даже если его появление не соответствует идее репрезентативности);

3) *достаточный объем* (если первые корпуса достигали миллиона слов, то объем современных корпусов исчисляется сотнями миллионов и миллиардами, например, объем корпуса английского языка Bank of English превышает 2,5 млрд слов);

4) *экономичность* (корпус текстов должен экономить усилия исследователя при изучении проблемной области, т.е. быть не просто строгим подмножеством текстов проблемной области, но, по возможности, быть наиболее «экономичным»);

5) *структуризация материала* (в корпусе должны быть выделены адекватные корпусу единицы хранения);

б) *компьютерная поддержка* (поддержка корпуса текстов ком-

плексом программ по обработке данных, обеспечивающих выявление контекстов слова, статистическую инвентаризацию, автоматическую словарную обработку и т.д.).

Лингвистическая разметка подразумевает присвоение словам особых кодов. Каждому коду соответствует определенный набор грамматических признаков, характеризующих данное слово. Коды также известны как тэги (от англ. tag – ярлык, метка), а сам процесс приписывания словам тэгов соответственно имеет название тэггинг (от англ. tagging).

Типы разметки, которые может содержать корпус, можно условно подразделить на лингвистические и внешне лингвистические (экстралингвистическими). К последним относятся:

- разметка, отражающая особенности форматирования текста (заголовки, абзацы, отступы и т.д.);

- разметка, касающаяся сведений об авторе и тексте. 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 и т.д. Это кодирование информации имеет название *метаразметка*.

Структурные метки несут информацию о статусе каждой единицы (глава, абзац, предложение, словоформа), а *собственно лингвистические* описывают лексические, грамматические и прочие характеристики элементов текста.

В соответствии с уровнем лингвистического описания различают морфологическую (определение части речи и морфологических категорий), синтаксическую (определение синтаксических связей), семантическую (категории, характеризующие значение слова), анафорическую (характеристика референтных связей, например, местоимений), просодическую (характеристика ударения и интонации), дискурсную (обозначение пауз, повторов, оговорок устной речи) и некоторые другие виды разметки.

В частности, предложение *Этой весной опять расцвела акация* может быть размечено следующим образом:

Этой — МЖЕТ21 весной — СЖЕТ22 опять — Н22 расцвела — ГЖЕП33 акация — СЖЕИЧ42

Первый индекс указывает на часть речи (М — местоимение, С — существительное, Н — наречие, Г — глагол), второй обозначает род, третий — число, четвертый — падеж или время (у глагола), первая цифра указывает на число слогов, а вторая — на ударный слог. Для разметки корпуса сообщений Твиттера при проведении международного исследовательского проекта по изучению данного жанра (Ганновер, 2010) были использованы, в частности, следующие виды меток:

<STDS> (стандартное написание), <KOKS> (использование только строчных букв), <KOGS> (использование только прописных букв), <GDOP> (удвоение графем), <GAUS> (выпадение графем), <GZUV> (написание лишней графемы) и т.д.

В зависимости от характера собранных в корпусе текстов, от их разметки и некоторых других факторов различают следующие виды корпусов (табл. 4).

Таблица 4

Классификация корпусов

№	Признак	Виды корпусов
1	Форма хранения	Звуковые, письменные, смешанные
2	Язык текстов	Русский, английский и т.д.
3	«Параллельность»	Одноязычные, двуязычные, многоязычные
4	Стиль	Литературные, диалектные, разговорные, публицистические, терминологические, смешанные
5	Способ доступа	свободно доступные, коммерческие, закрытые
6	Разметка	Размеченные, неразмеченные
7	Характер разметки	Морфологические, синтаксические, семантические, просодические и т.д.

Наиболее важным видом корпусов является универсальный национальный корпус, создаваемый для разных национальных языков. Создание и расширение универсальных национальных корпусов представляет собой одну из важнейших задач корпусной лингвистики.

Универсальный национальный корпус – это собрание текстов конкретного естественного языка представительное по отношению к всему языку, которое может служить для исследования самых разнообразных явлений этого языка.

Общепризнанный образец универсального национального корпуса – Британский национальный корпус (BNC) (www.natcorp.ox.ac.uk). Для русского языка таким представительным корпусом является Национальный корпус русского языка (НКРЯ) (<http://www.ruscorpora.ru>). Среди корпусов славянских языков выделяется Чешский национальный корпус (<http://ucnk.ff.cuni.cz>), созданный в Карловом университете Праги. Национальные корпуса существуют также для немецкого, китайского, финского и других языков.

Одним из первых известных корпусов является Брауновский корпус (Brown Corpus), созданный в 1963 г. в Брауновском университете (США) для построения частотного словаря американского варианта английского языка. Его объем составлял 1 млн слов. Создатели корпуса (У. Френсис и Г. Кучера) разработали строгую процедуру отбора текстов:

в корпус вошли 500 фрагментов прозаических текстов, созданных американскими авторами и напечатанных в 1961 г., по 2000 словоупотреблений каждый. Тексты представляли 15 наиболее распространенных жанров информативной и художественной прозы.

Поиск в корпусе в соответствии с запросом пользователя обеспечивается с помощью специальных программ – корпусных **менеджеров**. Корпусный менеджер (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. Результаты поиска обычно выдаются в виде конкорданса (поэтому корпусные менеджеры еще называют конкордансерами), где искомая единица представлена в ее контекстном окружении с представлением частотных характеристик отдельных языковых единиц, граммем и т.п.

Таким образом, корпус, представляющий собой размеченное собрание текстов с объемом слов не менее 100 млн, дает широкие возможности как для прикладных (работа над принципами автоматической разметки), так и для исследовательских целей.

Задание 2.

1. Выберите один из корпусов из списка ниже и охарактеризуйте его по следующим критериям: количество словоупотреблений, вид корпуса (по разным признакам).

- Британский национальный корпус (www.natcorp.ox.ac.uk),
- Американский национальный корпус (www.americannationalcorpus.org),
- Банк английского языка (Bank of English) (www.collins.co.uk/Corpus/CorpusSearch.aspx),
- Национальный корпус русского языка (www.ruscorgo.ru),
- Национальный корпус русского литературного языка (<http://www.narusco.ru/>),
- Компьютерный корпус текстов русских газет конца XX века (www.philol.msu.ru/~lex/corpus),
- Словарь-корпус языка А.С. Грибоедова (<http://febweb.ru/feb/concord/abc/>),
- Корпус института немецкого языка в Мангейме (www.ids-mannheim.de/kl/),
- Корпус испанского языка (исторический) (<http://www.corpusdelespanol.org/>),
- Корпус современного итальянского языка CORIS/CODIS (<http://www.cilta.unibo.it/ricerca.htm>),
- Польский национальный корпус (<http://korpus.ia.uni.lodz.pl/>).

2. Составьте глоссарий по теме «Корпусная лингвистика». Используйте для этого сетевые ресурсы. Включите в глоссарий определения

следующих понятий: конкорданс, рандомизация, коллокация, подмассив, парсинг, лемматизация, корпус-менеджер.

3. Найдите сетевые ресурсы по теме «корпусная лингвистика» и кратко охарактеризуйте их.

4. Откройте страницу Web-сайта Национального корпуса русского языка (РНК) (<http://www.ruscorpora.ru>).

5. На главной странице сайта выберите ссылку «Поиск в корпусе».

Откроется страница «Поиск в корпусе», который содержит литературную прозу (художественную и нехудожественную, письменную и устную) 19-20 веков. Именно этот корпус является основной частью Национального корпуса русского языка.

Если необходимо работать с поэтической, диалектной речью, или, например, с Параллельным корпусом, то необходимо осуществить дополнительные шаги:

а. На странице «Поиск в корпусе» выбирается ссылка «Задать подкорпус».

б. Откроется окно «Выбор подкорпуса».

Чтобы отобразить тексты определенного автора, надо в разделе «Основные параметры текста» в строке «Автор текста» набрать фамилию автора: Гоголь.

с. После этого пролистать эту страницу до конца, найти кнопку «Далее», щелкнуть по ней мышкой, появится список найденных текстов.

6. Откройте веб-страницу Корпуса русского литературного языка (КРЛЯ) (www.narusco.ru) и Британского национального корпуса (БНК) (www.natcorp.ox.ac.uk). Введите в строку поиска этик корпусов слово русский / Russian.

7. Заполните таблицу 5.

Таблица 5

	РНК	КРЛЯ	БНК
Количество вхождений			

8. Выпишите 3 любых контекста использования слова русский / Russian в трех рассмотренных корпусах. Укажите источник каждого примера в таблице 6.

Таблица 6

№ примера	РНК	КРЛЯ	БНК
1			
2			
3			

9. Сравните морфологические характеристики выписанных слов (существительное/прилагательное).

Таблица 7

№ примера	РНК	КРЛЯ	БНК
1			
2			
3			

10. Сравните значение выписанных слов. Для этого посетите веб-страницы толковых словарей www.gramota.ru/slovari и <http://oxforddictionaries.com>. Определите, в каком значении рассматриваемое слово встречается в контекстах. Впишите результат в таблицу 8.

Таблица 8.

№ примера	РНК	КРЛЯ	БНК
1			
2			
3			

11. К каким выводам вы пришли при сравнении морфологической и лексической характеристики одного и того же слова, включенного в разные корпуса?

12. Как можно использовать рассмотренные корпуса в лингвистическом исследовании?

К числу известных и наиболее часто используемых программ при разметке корпусов относятся такие программы как [AntConc](#), [WordSmith](#), [MonoConc Pro](#) и [CATMA](#).

AntConc (<http://www.antlab.sci.waseda.ac.jp/software.html>) является бесплатной, мультиплатформенной программой для проведения корпусных лингвистических исследований и управления данными. Она работает на любом компьютере под управлением Microsoft Windows, Linux. AntConc содержит семь инструментов, к которым можно получить доступ, нажав на клавишу табуляции в меню инструментов, или используя функциональные клавиши F1-F7.

Конкорданс (Concordance). Данный инструмент показывает результаты исследования формата KWIC (ключевое слово в контексте). Он позволяет увидеть, как слова и фразы обычно используются в разных контекстах.

График конкорданса (Concordance Plot). В этом инструменте все адреса для каждого элемента поиска представлены в виде “штрих-кода”, указывающего на место в файле, где находится элемент. График позволяет увидеть, какие файлы включают искомый элемент. Он также может быть использован для определения, где сталкиваются искомый элемент и кластер.

Просмотр файлов. В любое время целевой файл можно посмотреть в оригинальной форме, используя меню «просмотр файлов». Это позволяет более подробно исследовать результаты, полученные в других инструментах AntConc.

Кластеры (Clusters). Инструмент кластеры используется для создания упорядоченного списка кластеров, которые появляются вокруг поиска в целевом файле, перечисленные в левой части главного окна.

Расположение. Инструмент «расположение» показывает расположение элемента поиска. Это позволяет исследовать непоследовательные модели в языке.

Список слов (Word List). Данный инструмент подсчитывает все слова в корпусе и представляет их в упорядоченном списке. Это позволяет быстро найти, какие слова употребляются наиболее часто в корпусе.

Список ключевых слов (Keyword List). В дополнение к созданию списка слов, с помощью AntConc можно сравнить слова в целевом файле со словами, которые появляются в «базисном корпусе», чтобы создать список «Ключевых слов», которые являются наиболее частыми (или редкими) в целевых файлах.

Задание 3.

1. Запустите программу AntConc 3.3.5w.
2. Откройте в браузере страницу сложных английских текстов для чтения: <http://lengish.com/texts/category-2.html>.
3. Откройте текст [A cup of tea \(by Katherine Mansfield\)](#) и сохраните его как html файл.
4. Выполните команду File/Open files AntConc 3.3.5w и введите имя сохраненной веб-страницы.
5. Выберите инструмент **Concordance**.
6. Введите словоформу **could** в **Search Term** и нажмите **Start**, затем **could*** и нажмите **Start**.
7. Проанализируйте результаты поиска.

Работа с электронными словарями

Компьютерная лексикография представляет собой раздел прикладной лингвистики, нацеленный на создание компьютерных словарей, лингвистических баз данных и разработку программ поддержки лексикографических работ.

Основными задачами традиционной и компьютерной лексикографии являются определение структуры словаря и зон словарной статьи, а также разработка принципов составления различных видов словарей.

Словарь традиционно определяется как организованное собрание слов с комментариями, в которых описываются особенности струк-

туры и/или функционирования этих слов. *Электронный* (автоматический, компьютерный) *словарь* – это собрание слов в специальном компьютерном формате, предназначенное для использования человеком или являющееся составной частью более сложных компьютерных программ (например, систем машинного перевода). Соответственно, различаются *автоматические словари конечного пользователя-человека* (АСКП) и *автоматические словари для программ обработки текста* (АСПОТ).

Автоматические словари, предназначенные для конечного пользователя, чаще всего являются компьютерными версиями хорошо известных обычных словарей, например:

- Оксфордский словарь английского языка (www.oed.com),
- автоматический толковый словарь английского языка издательства «Коллинз» (www.mycobuild.com),
- автоматический вариант «Нового большого англо-русского словаря» под ред. Ю.Д. Апресяна и Э.М. Медниковой (<http://eng-rus.slovaronline.com/>),
- словарь Ожегова онлайн (<http://slovarozhegova.ru>).

Автоматические словари такого типа практически повторяют структуру словарной статьи обычных словарей, однако они обладают функциями, недоступными своим прототипам, например, осуществляют сортировку данных по полям словарной статьи (ср. отбор всех прилагательных), проводят автоматический поиск всех вокабул, имеющих в толковании определенный семантический компонент, и т.д.

Автоматические словари для систем машинного перевода, автоматического реферирования, информационного поиска и т.д. (АСПОТ) по интерфейсу и структуре словарной статьи существенно отличаются от АСКП. Особенности их структуры, сфера охвата словарного материала задаются теми программами, которые с ними взаимодействуют. Такой словарь может содержать от одной до сотни зон словарной статьи. Чрезвычайно разнообразны и области лексикографического описания: морфологическая, лексическая, синтаксическая, семантическая и т.д.

Структура традиционного словаря обычно включает следующие компоненты:

- введение, объясняющее принципы пользования словарем и дающее информацию о структуре словарной статьи;
- словник, включающий единицы словаря: морфемы, лексемы, словоформы или словосочетания; каждая такая единица с соответствующим комментарием представляет собой словарную статью;
- указатели (индексы);
- список источников;

- список условных сокращений и алфавит.

В электронных словарях из названных компонентов обязательным является *словник*, в онлайн-словарях имеется также *алфавит* с гиперссылками, ведущими к тексту словарной статьи. Практически в каждом электронном словаре, предлагаемом на диске (оффлайн-словарь) или в Интернете (онлайн-словарь) имеется функция (*автоматического поиска*, позволяющая значительно экономить усилия пользователя при работе со словарем).

Гиперссылки позволяют связывать разные словари друг с другом, так что в итоге онлайн- или оффлайн-словари оказываются коллекциями или порталами словарей. Получив необходимую информацию, например, о значении слова, пользователь одним нажатием ссылки может перейти к комментариям этого слова в других словарях и узнать особенности его толкования в специальных отраслях знания (терминологические словари) или получить дополнительную лингвистическую информацию о его форме.

Структура словарной статьи достаточно типична и обычно включает следующие зоны словарной статьи, актуальные как для традиционной, так и для компьютерной лексикографии:

- лексический вход (вокабула, лемма);
- зона грамматической информации;
- зона стилистических помет;
- зона значения;
- зона фразеологизмов;
- зона этимологии;
- зона примера и источника примера.

Можно выделить зоны словарной статьи, обязательные для всех словарных единиц, и факультативные зоны. Обязательной зоной словарной статьи для разных видов словарей является лишь лексический вход, все остальные зоны зависят от типа словаря: например, для толкового словаря необходима *зона значения*, а для орфоэпического она необязательна. Зона фразеологии отсутствует в комментариях слов, не используемых в устойчивых сочетаниях, а наличие зоны примера и его источника зависит от принципов, лежащих в основе создания словаря. Объем предлагаемой словарной информации должен соответствовать виду словаря.

Классификацию компьютерных словарей можно осуществлять на тех же принципах, что и классификацию обычных словарей. Традиционно выделяются лингвистические, энциклопедические, лингвострановедческие и терминологические словари. В *лингвистических словарях* описываются сами слова и их значения, особенности употребления, структурные свойства, сочетаемость, соотношение с лексическими системами других языков и т.д. В *энциклопедических словарях* описы-

ваются понятия, факты и реалии окружающего мира, т.е. экстралингвистическая информация.

Среди лингвистических словарей можно выделить несколько их видов:

- *толковые*, имеющие целью толкование (объяснение) значений слов и их употребления в речи, включающие дескриптивные и нормативные словари, которые, кроме того, могут быть общими и частными, среди последних выделяются, например, фразеологические словари, словари иностранных слов и т.д.;

- *словари-тезаурусы*, отличающиеся расположением словарной статьи, которое подчинено не алфавитному, а тематическому принципу, например, тезаурус русской идиоматики включает семантическое поле «УХОД, ОТЪЕЗД, БЕГСТВО», которое помещена в категорию «ДВИЖЕНИЕ», семантическое поле «ДАВНО» помещено в категорию «ВРЕМЯ» и т.д.;

- *двуязычные (переводные) словари*, например, «Англо-русский словарь» В.К. Мюллера (1-е издание появилось в 1943 г.), «Французско-русский словарь активного типа» под ред. В.Г. Гака и Ж. Триомфа и др.;

- *ассоциативные словари*, объектом которых является сфера ассоциативных отношений в лексике; словарная статья такого словаря включает лексему-стимул и список упорядоченных по частоте и алфавиту (с указанием частоты) реакций, полученных в психолингвистическом эксперименте, например: «Ассоциативный тезаурус современного русского языка»;

- *исторические и этимологические словари*, предоставляющие информацию об истории слов, начиная с определенной даты на протяжении некоторого периода, с указанием возникновения новых слов и значений, их отмирания и видоизменения, или объясняющие происхождение слов;

- *словари языковых форм*, которые фиксируют особенности формы слов и в которых толкования значений отсутствуют или играют вспомогательную роль, например, орфографические и орфоэпические, словообразовательные и морфемные (показывают, как слова складываются из морфем и инвентаризуют их), грамматические (информация по каждому слову), позволяющая построить любую грамматически правильную форму), обратные словари;

- *словари речевого употребления*: словари трудностей и сочетаемости слов;

- *ономастиконы*: антропонимические словари и топонимические словари;

- *нетрадиционные*, подвергающие словарному описанию нетипичные лингвистические объекты, например, «Словарь русских

политических метафор» А.Н. Баранова и Ю.Н. Караулова, словари поэтических метафор, эпитетов, авторские словари и словари конкордансов.

Например, известны такие электронные энциклопедии, как Энциклопедия Британника (www.britannica.com), «Большая энциклопедия Кирилла и Мефодия» (www.megabook.ru) и энциклопедия «Кругосвет» (www.krugosvet.ru).

Примерами переводных электронных словарей выступают АВВУУ Lingvo (www.lingvo.ru), TranslateIt! (www.translateit.ru) и Multitran(www.multitran.ru). Электронные толковые словари – это, в частности, словарь Merriam Webster (www.merriam-webster.com) и словарь французского языка «Trésor de la langue française» (<http://atilf.atilf.fr>). Формальными электронными словарями являются орфографические словари русского (<http://slovari.yandex.ru>) и английского (www.spellcheckonline.com) языков.

Большую коллекцию словарей разных видов на дисках и в Интернете предоставляет издательство Duden (немецкий язык, www.duden.de) и Larousse (французский язык, www.larousse.fr).

Компьютерные словари обычно создаются на базе корпусов текстов с использованием средств автоматической обработки и поиска словарных единиц. Для этого привлекаются специальные программы – базы данных, компьютерные картотеки, программы обработки текста, которые позволяют автоматически формировать словарные статьи, хранить словарную информацию и обрабатывать ее. Так, создание электронного словаря, согласно А.Н. Баранову, включает следующие этапы:

1. формирование корпуса текстов и параллельно создание словника;
2. автоматическое формирование корпуса примеров;
3. написание словарных статей;
4. ввод словарных статей в базу данных (БД);
5. редактирование словарных статей в БД;
6. корректура текста в БД;
7. порождение текста словаря и формирование оригинал-макета;
8. печать словаря.

Приведенное описание процесса создания электронного словаря может корректироваться в зависимости от его вида, исследовательских принципов и других факторов, но в любом случае использование компьютеров и уже готовых корпусов текстов в компьютерной лексикографии позволяет уменьшить количество этапов в процессе создания электронного словаря и сэкономить время практически на каждом из них.

Так, вместо создания словарной карточки в компьютерной лексикографии используются базы данных. Записи баз данных дают возможность автоматически сортировать массив по выбранным параметрам, отбирать нужные примеры, объединять их в группы и т.д. Специализированных программных оболочек для лексикографических целей на рынке практически нет. Для этих целей вполне подходят современные базы данных типа ACCESS или PARADOX. Для поиска примеров создатели словарей могут использовать компьютерные программы построения конкордансов, например, DIALEX. Для создания оригинал-макета (верстки) словарей привлекаются издательские системы типа PageMaker или WinWord, которые позволяют приписывать стили зонам словарных статей, алфавитизацию, создание указателей и т.д.

Единственным примером специализированной компьютерной программы, предназначенной для компьютерных лексикографических работ, является «Программа автоматизированного составления и обработки словников» (авторы: М.В. Литус, Е.В. Литус) [5].

Электронные словари имеют положительные стороны не только в процессе их создания, но и в процессе использования. В частности, выделяются следующие преимущества в использовании электронных словарей:

- электронные словари позволяют по-разному представить содержание словарной статьи, в том числе с помощью разнообразных графических и мультимедийных средств, которые не используются в обычных словарях;
- в выдаваемой информации находят отражение такие технологии как, морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п.;
- становится возможным быстро получить информацию, которая непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме;
- электронный словарь позволяет быстро реагировать на изменения в языке и мире.

Несмотря на наличие значительного числа преимуществ использования электронных словарей, остаются нерешенными некоторые проблемы, актуальные как для традиционной, так и для компьютерной лексикографии. В словарях должно найти отражение понятие лексической функции, позволяющее систематически описывать несвободную сочетаемость слов, иллюстрируемую следующими примерами русского языка: «войну ведут», а «экзамен – держат», «теории выдвигают», а «мысли подают» [4].

Задание 4.

1. Запустите программу *автоматизированного составления и обработки словарей* с преподавательского диска в архиве obr_slv.zip.
2. Прочитайте описание принципов работы с программой автоматизированного составления и обработки словарей: GetStart.doc.
3. Откройте текст [A cup of tea \(by Katherine Mansfield\)](#) и сохраните его как txt файл.
4. Выберите текстовый файл в программе и создайте словарь.
5. Выполните поиск слова again и просмотрите контекст.
6. Проанализируйте результаты поиска.

Задание 5.

1. Посетите сайт www.rvb.ru/soft/catalogue/index.html. В разделе 7 – словари, и тезаурусы выберите «Словарь сокращений русского языка». Протестируйте предлагаемый онлайн-словарь, введя любое сокращение русского языка. Представьте результаты работы в таблице 9.

Таблица 9

Введенное сокращение	Расшифровка сокращения
	1.
	2.
	3.

2. Посетите сайт www.merriam-webster.com. Введите слово culture в строку поиска. Определите зоны словарной статьи для этого слова в словаре Merriam Webster и представьте результаты вашего анализа в таблице 10.

Таблица 10

Зоны словарной статьи	Данные для слова culture в электронном словаре Merriam Webster

3. Посетите сайт www.ozhegov.org. Введите слово культура в строку поиска. Определите зоны словарной статьи для этого слова в электронной версии словаря Ожегова и представьте результаты вашего анализа в таблице 11.

Таблица 11

Зоны словарной статьи	Данные для слова culture в электронном словаре Ожегова

4. Сравните количество зон словарной статьи в двух рассмотренных словарях, в каком словаре их больше? Какие нужные, на ваш взгляд, зоны словарной статьи отсутствуют в рассмотренных словарях? С каким словарем вам было удобнее работать и почему?

Работа с терминологическими банками данных

Одним из перспективных направлений компьютерной лексикографии и прикладной лингвистики в целом является работа над электронными терминологическими словарями и банками данных. Построением специальных терминологических словарей занимается терминография, представляющая собой особый раздел лексикографии. В то же время терминография тесно связана с терминоведением – наукой о терминах. Соответственно, **компьютерная терминография** – это наука о составлении электронных терминологических словарей.

Принципы компьютерной терминографии в общем и целом то же, что и рассмотренные выше принципы компьютерной лексикографии. Их отличия связаны только с основным объектом словарного описания: в лексикографии это обычное слово или другие языковые единицы (морфема, словосочетание, предложение и т.п.), а в терминографии – термин.

Термин – это слово (словосочетание) метаязыка науки или области практической деятельности человека, имеющее четкое и (по возможности) однозначное определение, требующее специальных знаний из соответствующей профессиональной сферы. Так, слово «Интернет» для обычного человека выступает общеупотребительным,

а знакомство с соответствующим понятием ограничивается теми манипуляциями, которые человек производит с Интернетом (выбор провайдера услуг, тарифа, настройка подключения и некоторые другие). Для специалиста в компьютерных сетях это слово связано с огромным пластом предметного знания (история появления, технические характеристики, альтернативные Интернету виды связи и т.д.), соответственно, для специалиста оно выступает термином.

Из приведенных пояснений становится понятно, что понятие термина задается через его свойства, реализуемые в терминосистеме. *Терминосистема в целом отражает соответствующую область знания, а каждый ее компонент (термин) называет или характеризует составляющие этой области знания.*

Поскольку области знания объективны, а термины и терминосистемы «привязаны» к конкретному языку или даже к конкретной научной школе, важной задачей терминографии становится стандартизация и унификация терминов, а также их однозначный перевод на разные языки мира.

Унификации терминосистем служат терминологические стандарты. Но самих стандартов по организации терминосистем в мире сейчас более 20 тысяч; кроме того, существуют терминологические стандарты самых разных уровней: международного, государственного и даже уровня отдельных компаний и фирм.

Современные компьютерные технологии позволяют обрабатывать и сохранять большие массивы терминов по различным областям знания. Такие массивы терминов называются терминологическими базами (банками) данных (ТБД). По количеству задействованных в базе данных языков различаются переводческие (многоязычные) и информационно-нормативные (одноязычные) ТБД. Крупные ТБД имеются:

- в Научно-исследовательском институте комплексной информации по стандартизации и качеству (ВНИИКИ) (www.vniiki.ru);
- в Международной организации по стандартизации (англ. ISO = International Organization for Standardization, www.iso.org/obp/ui).

Кроме того, термины определенной предметной области собираются и описываются в словарях специальных терминов. Эти словари могут быть дескриптивными и нормативными, общими и частными, толковыми и переводными, алфавитными и тезаурусными.

Большинство электронных терминологических словарей носит дескриптивный характер и представляет термины отдельной отрасли знания. При этом востребованы и толковые (одноязычные), и переводные (двухязычные или многоязычные) словари. Разнообразные терминологические словари русского языка (анатомический, эконо-

мический, психологический и т.д.) представлены, в частности, на портале Gramota.ru (www.gramota.ru/slovari/online), а переводные терминологические словари, относящиеся к разным отраслям знания, можно найти по адресу www.diclib.com.

При описании термина важными оказываются следующие его свойства, сопоставимые с отдельными зонами словарной статьи:

- 1) семантика: связь термина с обозначаемым понятием;
- 2) словоизменение: особенности образования морфологических форм термина;
- 3) словообразование: включение термина в словообразовательное гнездо, установление связей между однокоренными словами (ср. прилагательные коммуникативный и коммуникационный, относящиеся к разным значениям термина «коммуникация»);
- 4) синтаксические связи: управление, сочетаемость с другими терминами и не терминами;
- 5) парадигматические связи в терминосистеме: синонимы, антонимы, гиперо-гипонимические связи, пересечения значения, терминологические ряды;
- 6) произношение;
- 7) примеры использования в контексте;
- 8) происхождение;
- 9) переводные эквиваленты.

Задание 6.

1. Откройте главную страницу Европейского интерактивного терминологического банка данных ИАТЕ (<http://iate.europa.eu>). Введите в строку поиска аббревиатуру NLP.

2. Выберите исходный язык (Source language) English, языки перевода (Target languages) – немецкий (de) и французский (fr). В дополнительных опциях выберите раздел 3236-Information technology and data processing.

3. В открывшемся окне нажмите на надпись «Полная информация» (Full entry) первого значения. Результаты поиска скопируйте в таблицу 12.

Таблица 12

Язык	Зоны словарной статьи			
	Definition	Term	Term	Abbreviation
en – English				
de – Deutsch				
fr – Français				

4. Как вы можете прокомментировать возможности данного терминологического банка данных? Для каких целей и кем он может быть использован?

5. Ознакомьтесь с двумя множествами терминов: прилагательными и существительными.

Таблица 13

Прилагательные	Существительные
информационный	ресурс
мультимедийный	технология
цифровой	средства
электронный	платформа

6. Скомбинируйте перечисленные выше существительные и прилагательные с целью создания терминологических сочетаний, например: информационная платформа. Перечислите все получившиеся терминологические словосочетания в таблице.

Таблица 14

Термин	Словосочетания с данным термином
ресурс	
технология	
средства	
платформа	

7. С помощью систем поиска (google.ru, yandex.ru и т.п.) напишите словарную статью для одного из получившихся терминов по вашему выбору. Статья должна включать следующие обязательные зоны: лексический вход, определение, примеры использования, источники. Кроме того, включите в описание термина еще две зоны словарной статьи на ваш выбор. Результат внесите в таблицу 15.

Таблица 15

Зоны словарной статьи	Описание
Лексический вход	
Определение	
Примеры	
Источники	

Машинный перевод

Вопросы машинного перевода составляют одну из центральных областей использования информационных технологий в лингвистике и филологии. Это обусловлено, прежде всего, постоянно возрастающей практической потребностью современного общества в переводе значительного количества текстов различной функциональной направленности.

Доказательством возрастания потребности в переводе служат документы международных организаций, которые в обязательном порядке переводятся на языки стран-участников и обеспечивают рабо-

той большое количество профессиональных переводчиков, работа которых достаточно дорогая, и к тому же медленная. Так, нормой научно-технического перевода считается время 10 дней на авторский лист (24 страницы машинописного текста). Система машинного перевода позволяет получить перевод сотен авторских листов за 1 час. Кроме того, появляются новые области применения машинного перевода, например, тексты Интернета. По подсчетам исследователей, в Интернете встроенными системами перевода (SYSTRAN, TRADOS и STeam Translation) и сетевыми онлайн-словарями ежедневно выполняется 1 млн запросов на перевод текстов в различных форматах.

Системы машинного перевода моделируют работу человека-переводчика. Таким образом, суть машинного перевода та же, что и в случае его выполнения человеком, с той лишь разницей, что в этом процессе используются компьютеры. **Машинный (или автоматический) перевод (МП)** – выполняемое компьютером действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

С точки зрения роли человека в процессе выполнения МП различают следующие его виды:

МАНТ (Machine-assisted human translation) – перевод, осуществляемый человеком с использованием компьютера;

НАМТ (Human-assisted machine translation) – машинный перевод при участии человека;

ФАМТ (Fully-automated machine translation) – полностью автоматизированный машинный перевод.

В первом случае человек использует компьютерные инструменты, направленные на ускорение и упрощение процесса перевода, но собственно перевод текста выполняет сам человек. Вспомогательными системами компьютерной поддержки перевода здесь выступают электронные словари, терминологические базы данных.

Второй тип систем МП является своего рода промежуточным: здесь одинаково важно участие в процессе перевода и человека, и машины. В машину вводятся электронные словари, морфологические справочники и задается определенный алгоритм выполнения задачи перевода. Роль человека здесь сводится к выбору предлагаемых машиной решений и редактированию текста перевода.

Весьма наглядно такой тип систем МП иллюстрируется системами переводческой памяти (Translation Memory, ТМ). Идея таких систем заключается в хранении базы данных переводов, сделанных профессиональным переводчиком, для того чтобы в процессе перевода предлагать человеку уже готовый перевод фразы или фрагмента текста, если он уже был однажды переведен. ТМ-программы значительно повышают эффективность работы переводчика, избавляя его от рутинной, повторяющейся работы. Во многих фирмах, занимающихся переводом, владение одной из таких программ является существенным критерием при приеме на работу.

Эффективность полностью автоматизированных систем МП зависит от того, в какой степени в них учитываются объективные законы функционирования языка и мышления. Но эти законы пока еще недостаточно изучены, и перед создателями систем МП возникает множество проблем, отражающихся в недостаточном качестве результата МП. По мере усложнения систем МП и включения в них новых этапов автоматического анализа и синтеза текста выделяют три поколения таких систем:

1. П-системы системы прямого перевода (*direct systems*);
2. Т-системы системы с синтаксическим преобразованием исходного текста (от англ. *transfer* – преобразование);
3. И-системы – системы с семантическим и прагматическим анализом (*interlingua* язык-посредник).

Первый тип систем МП (П-системы) включает лишь этапы морфологического анализа и синтеза, поэтому результат работы таких систем представляет своего рода подстрочный перевод. Во втором типе систем МП (Т-системах) привлекаются методы синтаксического анализа и синтеза, причем в зависимости от их уровня (поверхностный, глубинный или синтактико-семантический) выделяют и разные виды Т-систем. Наиболее сложный тип систем МП – И-системы – включает наряду с лингвистической и экстралингвистическую информацию, т.е. семантику и прагматику предметной области. Поэтому после этапов морфологического и синтаксического анализа фразы исходного текста алгоритм И-системы включает этап семантического анализа. Его результатом служат семантические представления фраз ИЯ и ПЯ, обеспечивающие эквивалентность их смысла.

В итоге в целом схема машинного перевода включает следующие этапы:

1. ввод в компьютер текста на ИЯ;
2. его морфологический анализ, т.е. определения части речи и морфологических характеристик каждого слова;
3. синтаксический анализ каждого предложения текста ИЯ (поиск основных членов предложения и определение типов синтаксических связей между ними, выражаемых в виде дерева зависимостей или дерева непосредственных составляющих);
4. семантический анализ каждого предложения ИЯ, в результате которого создается семантическое представление этого предложения, независимое от типа языка (общее и для ИЯ, и для ПЯ);
5. синтаксический синтез предложений ПЯ (создание предложений правильной синтаксической структуры, соответствующей правилам ПЯ и типу синтаксической структуры предложения на ИЯ);
6. морфологический синтез каждого слова в составе отдельных предложений текста ПЯ (постановка слов ПЯ в нужных морфологических формах);
7. вывод текста на ПЯ.

Отдельные трудности процесса МП связаны с необходимостью определения анафорических связей в текстовом целом (anaphora resolution), снятия омонимии на разных уровнях, а также с необходимостью привлечения в процесс перевода экстралингвистических знаний.

Важность анафорических связей определяется достаточно активным использованием в тексте языковых выражений, которые не могут быть поняты без обращения к предыдущему контексту. Такими выражениями выступают, к примеру, анафорические местоимения он или he. Установление того, к какому языковому выражению из предыдущего текста относится анафорическое местоимение и к какой сущности реального мира (референту) местоимение и его антецедент отсылает, важно как для понимания всего текста, так и для правильного построения синтаксического и морфологического представления текста. Правильная интерпретация анафорического местоимения требует привлечения данных всех языковых уровней, выхода за рамки одного предложения и привлечения прагматического анализа всего текста.

В итоге для функционирования систем МП требуется *лингвистическое, программное и информационное обеспечение* систем МП. *Лингвистическим обеспечением* таких систем выступают словари слов и словосочетаний с соответствующими признаками для ИЯ и ПЯ; морфологические таблицы суффиксов и окончаний для ИЯ и ПЯ; базы грамматических правил и др. *К программному обеспечению* относятся программы выполнения перевода, ведения словарей, формирования базы правил и т.д. *Информационное обеспечение* – представляет база экстралингвистических знаний о предметной области.

К числу наиболее распространенных систем МП относятся:

- Stylus система МП, включающая множество словарей по разным предметным областям;
- Universal Translator – многоязычная система МП;
- Polyglossum – многоязычная система МП с широким набором предметных словарей;
- Promt многоязычная система МП, содержащая множество словарей по разным предметным областям;
- WebTranSite – система для перевода веб-страниц (сам процесс перевода веб-страниц и сообщений компьютерных программ называется локализацией).

Сравнение и оценка систем МП осуществляется по следующим параметрам (Framework for the Evaluation of Machine Translation, FEMT):

- характеристики программного обеспечения: надежность системы, удобство использования, скорость работы, возможность обновлений, эффективность, мобильность и т.п.;
- характеристики пользователя и задач перевода: особенности пользователя, автора и текста, а также назначение перевода;
- особенности системы МП: стратегия построения системы, лингвистические ресурсы и т.п.;

- специфика выходного текста: точность, целостность, стиль и т.п., а также наличие ошибок любого характера.

В частности, системы МП письменных текстов в значительной степени отличаются от систем перевода устной речи как по программному обеспечению (в последнем случае обязательно включение в процесс МП этапов автоматического анализа и синтеза устной речи), так и по тематике. Системы для перевода устного диалога обычно ориентированы на узкую тематику: резервирование мест в гостинице, определение маршрута проезда по городу и т.д. Соответственно, и оценку каждой из систем МП нужно производить с учетом их названных особенностей.

Задание 7.

1. Протестируйте работу разных систем МП, размещенных в Интернете (www.translate.ru от компании Promt и <http://translate.google.ru> от Google).

Для этого выполните автоматический перевод одного и того же текста (объем – 1-2 абзаца, ИЯ – русский, ПЯ – на ваш выбор, тематика – общая). Введите получившийся результат в таблицу 16.

Таблица 16

Исходный текст	Перевод 1, www.translate.ru	Перевод 2, http://translate.google.ru

2. Охарактеризуйте протестированные онлайн-переводчики по следующим параметрам:

Таблица 17

Критерий	Перевод 1	Перевод 2
Затраты времени на выполнение перевода		
Необходимость специальной подготовки пользователя (компьютерные, языковые знания и т.п.)		
Качество перевода (целостность текста, стилистическая однородность, наличие ошибок и т.п.)		
Необходимость постредактирования		

3. Отредактируйте один из вариантов перевода (Перевод 1 или Перевод 2).

Проанализируйте объем своей работы и заполните таблицу 18, характеризующую редактирование. При необходимости дополните таблицу собственными параметрами.

Таблица 18

Тип редактирования	Частота
Лексические замены переводов отдельных слов	
Удаление вариантов переводов	
Лексические замены переводов словосочетаний	
Исправление неверного согласования	
Исправление неверного управления	
Вставка дополнительных слов	
Вставка дополнительных словосочетаний	
Удаление лишних слов	
Изменение структуры предложения	

Прокомментируйте получившиеся результаты: какой вид редакторских работ востребован чаще всего, какой является самым сложным?

4. Сравните результаты перевода текстов разной функциональной принадлежности (темы), выполненного в онлайн-переводчике www.translate.ru.

Для этого наберите или скопируйте предлагаемые ниже фрагменты текстов в окно ввода, выберите в верхнем меню соответствующую тему, языки перевода (английский → русский) и нажмите «Перевести». Прокомментируйте, какие недостатки содержит результат перевода, внеся ваши комментарии в таблицу.

1) Техника: Компьютеры

Despite big changes in technology over the past couple of decades, IT departments and the duties of their staff have stayed pretty consistent. The classic model involves helpdesk agents, desktop support staff, systems and network administrators, DBAs and developers, and managers at various levels reporting to a CIO or technology director.

(Faas R. How Mobile, BYOD and Younger Workers Are Reinventing IT // PC World. 24.02.2012. www.pcworld.com).

2) Бизнес

In the early days of starting a business, you might be tempted to gloss over ownership structure, equity stakes, and other seemingly boring details. After all, you might think, as long as you keep taxes low, paperwork

uncomplicated, and partners motivated, better to deal with the big stuff first. But these decisions can have a significant cost down the road, particularly for entrepreneurs who seek outside investors.

(Mehta M. Structuring a Business with Investors in Mind // BusinessWeek. 22.02.2012. www.businessweek.com)

3) Прочее: Здоровье

Data from more than 250,000 men and women in 18 cohort studies were used to calculate the Lifetime risk of cardiovascular events, stratified according to risk-factor burden, with adjustment for the competing risk of death from noncardiovascular causes.

(Berry J.D. et al. Lifetime Risks of Cardiovascular Disease // The New England Journal of Medicine. 26.01.2012 www.nejm.org)

Таблица 19

Тема	Комментарии
1. Компьютеры	
2. Бизнес	
3. Здоровье	

5. Выполните перевод <http://translate.google.ru> на белорусский язык главной страницы сайта университета и сохраните в файле Perevod.html.

Задание 8.

1. Определите, какие из перечисленных веб-ресурсов не являются порталами:

Таблица 20

Ресурс	Да/нет	Обоснование
www.all-abc.ru		
www.gramota.ru		
http://pearsonpte.com		~
http://deutsche-sprache.ru		
www.english.language.ru		

2. Найдите с помощью различных поисковых систем и укажите в таблице по два примера русскоязычных и иноязычных Интернет-ресурсов (на английском, русском, немецком или французском языке).

Таблица 21

Вид ресурса	Русский	Иностр. язык
Электронная библиотека		
Электронный журнал		
Веб-квест		

3. Завершите работу с Mozilla Firefox.

Тема работы: Проектирование и создание словаря морфем английского языка.

Цель работы: Изучить процесс проектирования и создания базы данных на примере построения информационно-логической модели словаря морфем английского языка.

Ход работы

1. Открыть описание работы **Словарь морфем**.
2. Выполнить работу в группе проекта.
3. Отправить на проверку преподавателю.
4. Ответить на контрольные вопросы (устно на занятии).

Описание работы

Перед выполнением заданий прочитайте основные теоретические сведения.

Теоретические сведения

Проектирование информационно-логической модели для словаря методом «сущность-связь».

Предварительный этап проектирования инфологической модели предусматривает выполнение системного анализа и словесного описания информационных объектов предметной области. На первом этапе проектирования создается концептуальная схема базы данных (БД), которая затем преобразуется к реляционной схеме. В результате создается реляционная база данных. Рассмотрим одну из наиболее распространенных для формализованного представления инфологической модели предметной области моделей данных – модель “сущность-связь”. Модель “сущность-связь” разработана П. Ченом в 1976 году. Описание предметной области осуществляется в виде схем (диаграмм), на которых с помощью графических объектов представлены информационные объекты (сущности), описывающие их реквизиты (атрибуты) и связи между ними. Достоинством такого подхода является наглядность получаемых моделей и возможность формализации всего процесса построения информационных моделей.

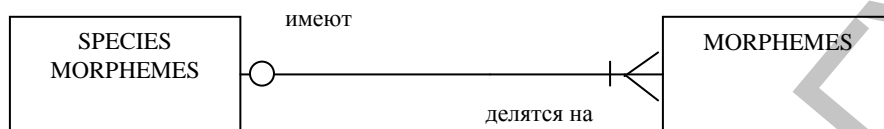
Основные понятия модели: сущность, связь и атрибут.

Сущность – это реальный или представляемый объект, информация о котором должна сохраняться и быть доступна. В диаграммах модели сущность представляется в виде прямоугольника, содержащего имя сущности. Каждый элемент сущности должен быть отличим от любого другого элемента этой же сущности.

Связь – это графически изображаемая ассоциация, устанавливаемая между двумя сущностями. В любой связи выделяются 2 конца, на каждом из которых указывается имя конца связи (в виде глагола), степень конца связи (сколько элементов данной сущности связывается) и обязательность связи (то есть, любой ли элемент данной сущности должен участвовать в этой связи). Связь представляется в виде

линии, соединяющей 2 сущности, при этом в месте соприкосновения связи с сущностью используется множественный вход в прямоугольник, если для связи могут использовать несколько элементов, и единственный – если в связи может участвовать только один элемент сущности. Обязательность связи изображается перпендикуляром, а необязательность – окружностью.

Примеры:



Конец связи **имеет**, означает, что один вид морфем имеет несколько конкретных морфем, например, к *root* – -estim-, -graph, therm-. Конец связи с именем **делятся на** означает, что морфемы делятся на виды и каждая конкретная морфема относится к одному виду. Трактовка изображенной диаграммы следующая: вид морфемы не обязательно имеет морфему, каждая морфема обязательно относится к виду. **Атрибутом** сущности является любой элемент, который служит для уточнения, идентификации, классификации, числовой характеристики или выражения состояния сущности. Имена атрибутов заносятся в прямоугольник, записываются под именем сущности, возможно с примерами.

Некоторый набор атрибутов назначается уникальным идентификатором (**ключом**).

Словесная характеристика предметной области.

Морфема – это минимальная значимая языковая единица, образуемая фонемами и входящая в состав слова. Морфемы формируют слова, вне слов они не существуют. Изучая морфему, мы фактически изучаем слово: его строение, его функции, то, как слово входит в речь.

В традиционной грамматике морфемы разграничиваются на основании двух критериев: позиционного критерия – положения морфем в слове относительно друг друга, и семантического (или функционального) критерия – вклада морфемы в совокупное значение слова.

В соответствии с этими критериями морфемы подразделяются на корневые морфемы (корни) и аффиксальные морфемы (аффиксы). Корневые морфемы выражают наиболее конкретную, «вещественную» часть значения слова и занимают в нем центральное положение. Аффиксальные морфемы выражают уточняющую, «спецификационную» часть значения слова: они уточняют, или трансформируют значение корня. Аффиксальная спецификация может быть двух видов: лексическая и грамматическая; поэтому по семантическому критерию аффиксы далее разделяются на лексические, или словообразователь-

ные (деривационные) аффиксы, которые вместе с корнем образуют основу слова, и грамматические, или словоизменительные аффиксы, выражающие значения различных морфологических категорий, таких как число, падеж, время и т.д. С помощью лексических аффиксов образуются новые слова; с помощью грамматических аффиксов изменяется форма одного слова [2].

По позиционному критерию аффиксы подразделяются на префиксы, которые располагаются перед корнем слова, например: underestimate, и суффиксы, которые располагаются после корня слова, например: underestim-ate. Префиксы в английском языке бывают только лексическими: так, слово underestimate образовано от слова estimate с помощью префикса under-. Суффиксы в английском языке могут быть либо лексическими, либо грамматическими; например, в слове underestimates суффикс -ate является лексическим, поскольку глагол estimate (v) образован от существительного esteem (n), а суффикс -s является грамматическим: он изменяет форму глагола to underestimate, выражая грамматическое значение 3-его лица единственного числа. Грамматические суффиксы также называются флексиями (окончаниями).

Грамматические суффиксы в английском языке обладают рядом особенностей, которые отличают их от флексий в других языках: поскольку грамматические суффиксы в английском языке представляются собой остатки старой флективной системы, их в английском языке немного, всего шесть: -(e)s, -ed, -ing, -er, -est, -en; большинство из них омонимично; так, -(e)s используется для обозначения множественного числа существительных (dogs), родительного падежа существительных (my friend's) и 3 лица ед. числа глагола (works); некоторые из них утратили флективные признаки и могут присоединяться к единицам больше, чем слово, например: his daughter Mary's arrival. Поэтому термин «флексия» редко используется для обозначения грамматических показателей слов в английском языке.

Морфемная структура слова может быть отражена линейно; например, структура слова underestimates может быть показана следующим образом: $W = \{[Pr + (R+L)] + Gr\}$, где W – слово, R – корень, L – лексический суффикс, Pr – префикс, Gr – грамматический суффикс.

В базе данных (словаре) необходимо хранить и обрабатывать информацию по словоформам, морфемам и их видам. В результате предпроектного обследования был определен перечень тех реквизитов, которые необходимо хранить в базе данных: код словоформы, словоформа на английском языке (underestimates...), транскрипция, часть речи, число (единственное, множественное), перевод на русский язык, примеры употребления, код вида морфемы, обозначение (R, L Pr Gr), класс морфем (root-morphemes (roots) and affixal morphemes (af-

fixes)), перевод класса морфем на русский язык (корневые морфемы (корни) и аффиксальные морфемы (аффиксы)), вид морфемы (*root, prefix, lexical suffix, inflexion (inflection) or grammatical suffix*), перевод вида морфемы на русский язык (корень, префикс, суффикс, флексия или грамматический суффикс), код морфемы конкретной словоформы, название (under- -estim- -ate- -s...).

Сформулированы следующие основные условия: одному виду морфемы в английском языке можно сопоставить несколько морфем, например вид «грамматический суффикс» один, а морфем относящихся к нему шесть: -(e)s, -ed, -ing, -er, -est, -en. В структуре каждой словоформы может быть один и несколько видовых морфем.

Создание информационно-логической модели.

Создание информационно-логической модели начинается с анализа взаимосвязей между атрибутами, выявления сущностей и определения ключей. Анализ взаимосвязей между атрибутами позволяет установить, что каждому коду словоформы соответствует конкретная словоформа на английском языке, транскрипция и часть речи. Для кода вида морфемы можно установить взаимооднозначное соответствие с обозначением и видом морфем. Аналогично коду морфемы конкретной словоформы соответствует название морфемы. Таким образом, можно выделить три сущности и назвать их **WORDS, MORPHEMES, SPECIES MORPHEMES**.

Для каждой сущности выделяются ключи, т.е. те атрибуты, которые однозначно идентифицируют записи. Например, для сущности **WORDS** уникальными (неповторяющимися) являются атрибуты: *Код_словоформы* и *Словоформа на английском языке*. И тот, и другой атрибут однозначно идентифицирует конкретную словоформу, который может быть выбран в качестве ключа. Однако в целях оптимизации размера базы данных и удобства работы с ней в качестве ключа обычно выбираются кодовые атрибуты.

Таким образом, структурирование данных предметной области позволило выделить три сущности и описывающие их атрибуты:

Далее необходимо установить взаимосвязи между сущностями, что осуществляется путем анализа типов связей между ключами с учетом сформулированных ранее условий описания предметной области. Связь между сущностями **SPECIES MORPHEMES** и **MORPHEMES** имеет тип «один-ко-многим». Это следует из того, что по условию один вид морфем имеет несколько конкретных морфем, но каждая конкретная морфема относится к одному виду. Между сущностями **WORDS** и **MORPHEMES** имеет место тип связи «многие-ко-многим», так как по условию каждая словоформа имеет от одной до нескольких морфем и каждая конкретная морфема принадлежит одной или нескольким словоформам (рис. 1).

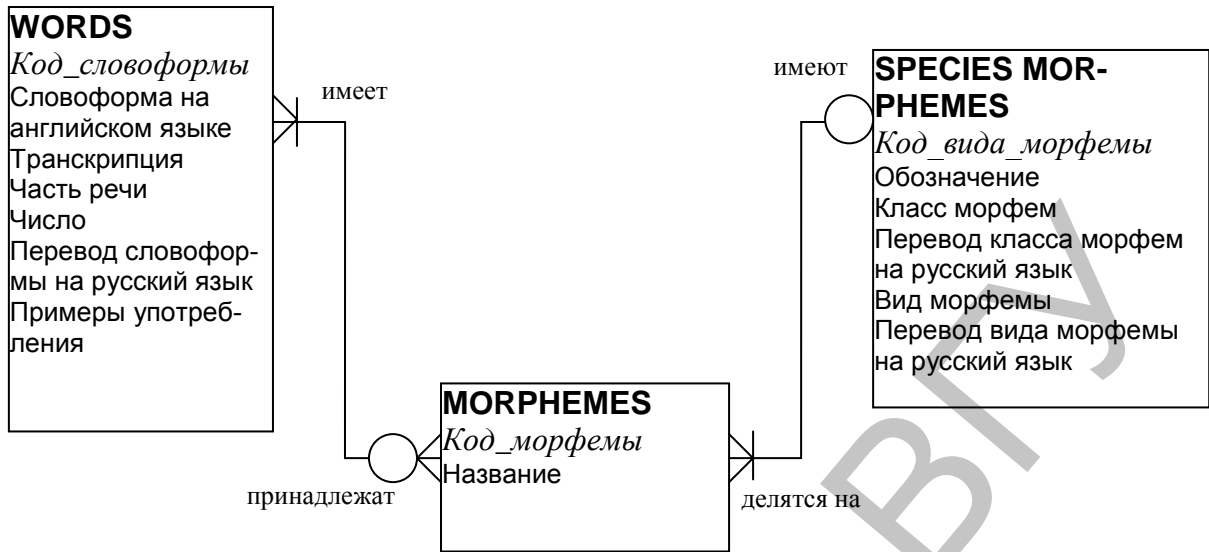


Рис. 1. Инфологическая модель «Dictionary of morphemes»

Далее осуществляется преобразование инфологической модели в реляционную модель (рис. 2).

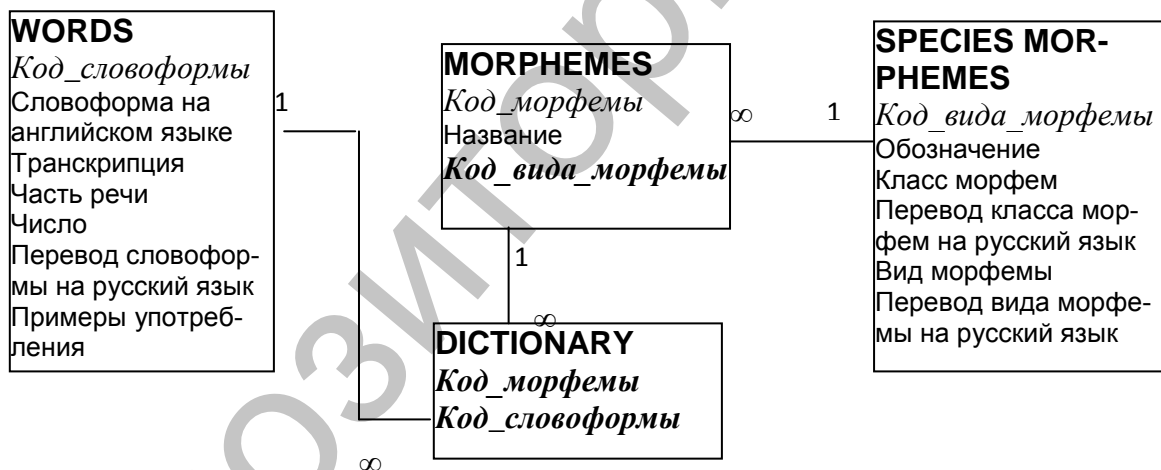


Рис. 2. Реляционная модель «Dictionary of morphemes»

Создание таблиц базы данных «Словарь морфем английского языка»

Для создания базы данных «Словарь морфем английского языка» в главном меню Access в пункте **Файл** выберите пункт **Создать**, затем в диалоговом окне Создание файла – пункт Новая база данных..., потом указать свой диск и папку, в которой будет храниться файл БД, затем введите имя файла «Морфемы» и нажмите кнопку Создать. На экране появится диалоговое окно (рис.3).

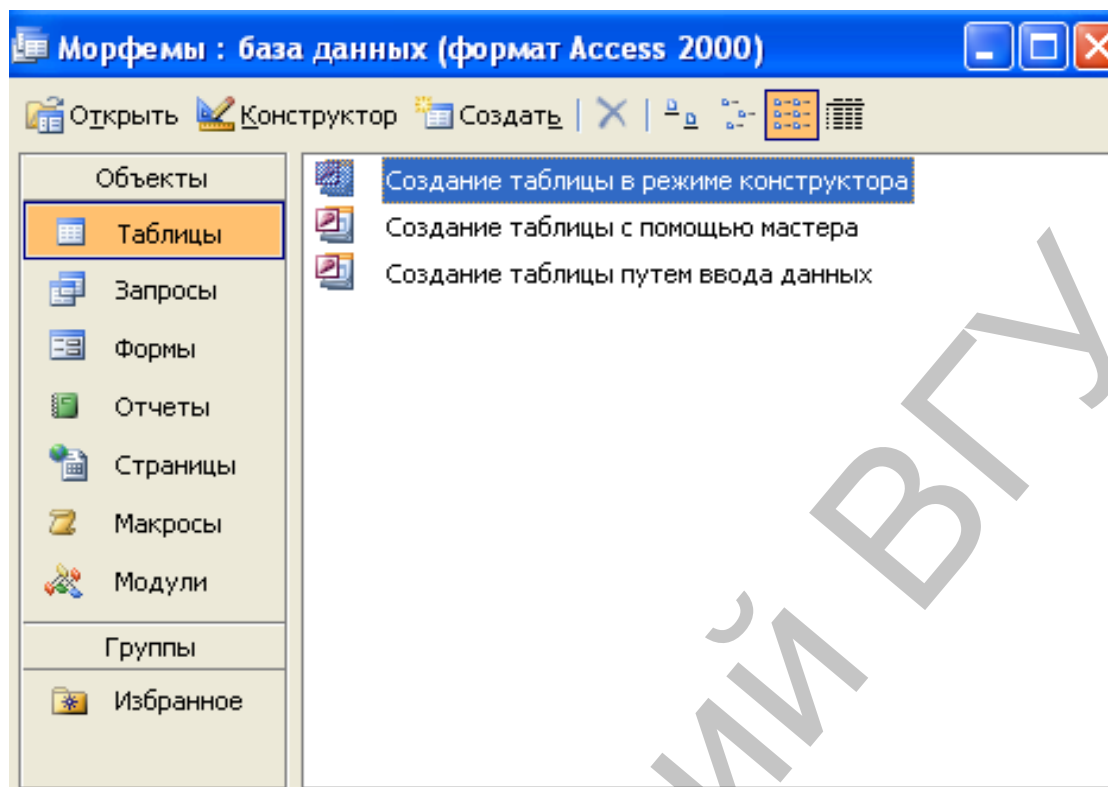


Рис. 3. Создание таблицы Microsoft Access

Для создания таблиц БД выберите Конструктор. При его использовании на экране появится окно, в котором необходимо описать структуру каждой таблицы базы данных, указав имя поля, его тип и при необходимости другие параметры (рис. 4).

Задание 1. Создайте структуры таблиц БД «Словарь морфем английского языка» SPECIES MORPHEMES, WORDS, MORPHEMES и DICTIONARY соблюдая последовательность и сохраните их.

Типы данных и свойства полей для таблицы SPECIES MORPHEMES:

Имя поля Код_вида_морфемы *Тип данных* Числовое *Размер поля* Длинное целое *Индексированное поле* Да (Совпадения не допускаются) **Ключевое поле.**

Имя поля Обозначение *Тип данных* текстовый *Размер поля* 2 **Обязательное поле** Да.

Имя поля Класс морфем *Тип данных* текстовый *Размер поля* 7 **Обязательное поле** Да.

Имя поля Перевод класса морфем на русский язык *Тип данных* текстовый *Размер поля* 7 **Обязательное поле** Да.

Имя поля Вид морфемы *Тип данных* текстовый *Размер поля* 18 **Обязательное поле** Да.

Имя поля Перевод вида морфемы на русский язык *Тип данных* текстовый *Размер поля* 22 **Обязательное поле** Да.

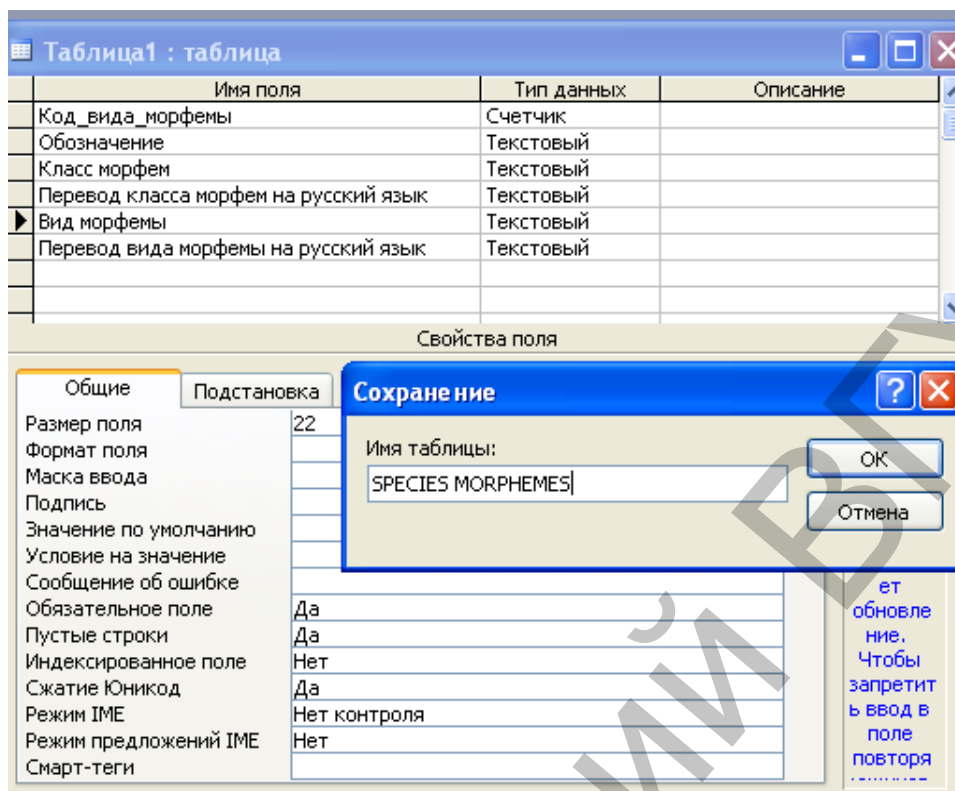


Рис. 4. Структура таблицы SPECIES MORPHEMES в режиме Конструктор

Типы данных и свойства полей для таблицы WORDS:

Имя поля Код_словоформы Тип данных Счетчик Размер поля Длинное целое Индексированное поле Да (Совпадения не допускаются) **Ключевое поле**.

Имя поля Словоформа на английском языке Тип данных текстовый Размер поля 50 Обязательное поле Да.

Имя поля Транскрипция Тип данных текстовый Размер поля 20 Обязательное поле Нет.

Имя поля Часть речи Тип данных текстовый Размер поля 20 Обязательное поле Да.

Имя поля Число Тип данных текстовый Размер поля 15 Обязательное поле Да.

Имя поля Перевод на русский язык Тип данных текстовый Размер поля 50 Обязательное поле Да.

Имя поля Примеры употребления Тип данных текстовый Размер поля 150 Обязательное поле Нет.

Типы данных и свойства полей для таблицы MORPHEMES:

Имя поля Код_морфемы Тип данных Счетчик Индексированное поле Да (Совпадения не допускаются) **Ключевое поле**.

Имя поля Название Тип данных текстовый Размер поля 20 Обязательное поле Да.

Имя поля Код_вида_морфемы Тип данных Числовой Размер поля Длинное целое Обязательное поле Да.

Типы данных и свойства полей для таблицы DICTIONARY:


Имя поля Код_морфемы Тип данных Числовой Размер поля Длинное целое Индексированное поле Да (Совпадения допускаются) **Ключевое поле**.

Имя поля Код_словоформы Тип данных Числовой Размер поля Длинное целое Индексированное поле Да (Совпадения допускаются) Ключевое поле.

После создания структуры таблиц необходимо разработать схему данных.

Схема данных – это графическое изображение взаимосвязей реляционных таблиц. Она позволяет наглядно показать структурную схему всей базы данных, а также обеспечить защиту от случайного удаления или изменения связанных данных. Взаимосвязь таблиц используется при создании запросов к БД, составных (подчиненных) форм, отчетов.

Создать схему данных можно двумя способами:

1. При помощи меню Microsoft Access: *Сервис/Схема данных*.
2. При помощи кнопок панели управления: щелкнуть мышью по кнопке *Схема данных* .

Чтобы можно было изменять и/или удалять записи в связанных таблицах, сохраняя при этом целостность данных, в Microsoft Access применяется каскадирование. С этой целью следует установить флажки *Каскадное обновление связанных полей* и *Каскадное удаление связанных полей*. Если установлен флажок *Каскадное обновление связанных полей*, то при изменении ключевого поля главной таблицы автоматически изменяются и соответствующие значения связанных записей. Если установлен флажок *Каскадное удаление связанных полей*, то при удалении записи в главной таблице удаляются и все связанные записи в подчиненной. Схема БД является отображением инфологической модели предметной области.

Задание 2. Разработайте схему данных для БД “Словарь морфем английского языка” (рис. 5).

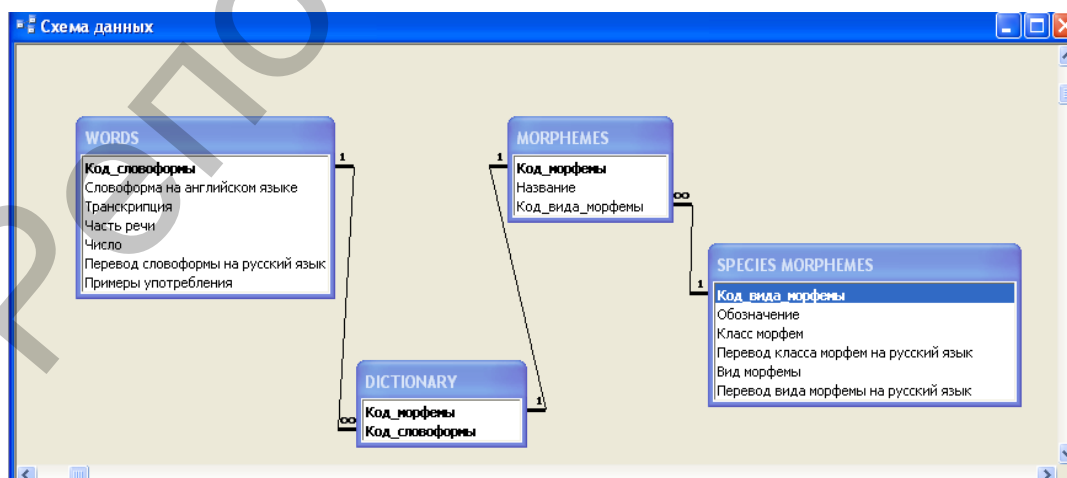


Рис. 5. Схема данных

Ввод данных в таблицы базы данных «Словарь морфем английского языка»

Ввод данных может осуществляться либо непосредственно в таблицы, либо через заранее созданные формы. Обычно ввод с помощью форм целесообразен тогда, когда данные вводятся в связанные таблицы или когда данные находятся в различных первичных документах. Перед вводом данных в связанные таблицы необходимо создать схему данных и в окне *Изменение связей* установить флажок *Обеспечение целостности данных*. Наличие схемы данных обязательно для построения составных связанных форм.

В Microsoft Access различают *простые* и *составные* (сложные) формы. Простые формы строятся на основе одной таблицы, а составные – нескольких таблиц.

Простая форма в Microsoft Access обычно представлена в одном из видов: *столбец*, *ленточная*, *табличная* (рис.5).

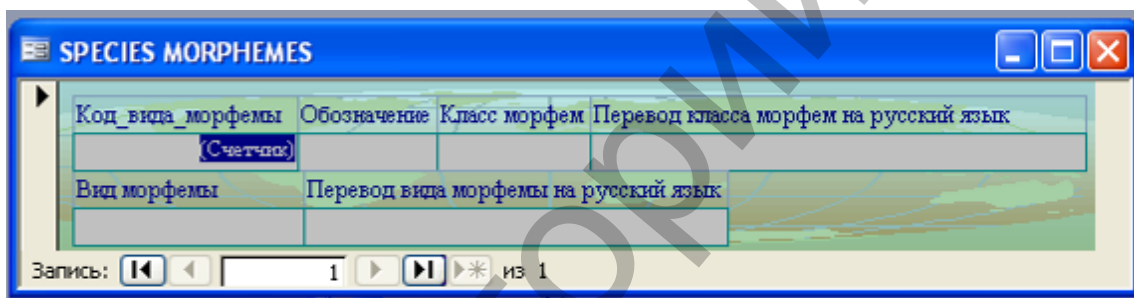


Рис. 6. Простая форма выровненная

Задание 3. Создайте простую форму в одном из видов: столбец, ленточная или табличная для ввода данных в таблицу SPECIES MORPHEMES.

Составные формы могут быть представлены в одном из двух видов: *подчиненная*, *связанная*. Подчиненная форма имеет иерархический вид, отображающий структуру первичного документа. Составная форма состоит из главной формы и одной или нескольких подчиненных форм. Вверху формы содержатся наименования и значения полей, входящих в главную форму, соответствующие общей части документа, а внизу формы отображаются наименования и значения полей, входящих в подчиненные формы (рис.7).

Таблицы, входящие в составную форму, должны быть связаны в отношении «один-ко-многим» или «один-к-одному». Одна из таблиц при проектировании формы объявляется главной. Если между таблицами имеется связь «один-ко-многим», то в главную форму входят поля главной таблицы и могут входить поля таблиц, связанные с главной со стороны «много».

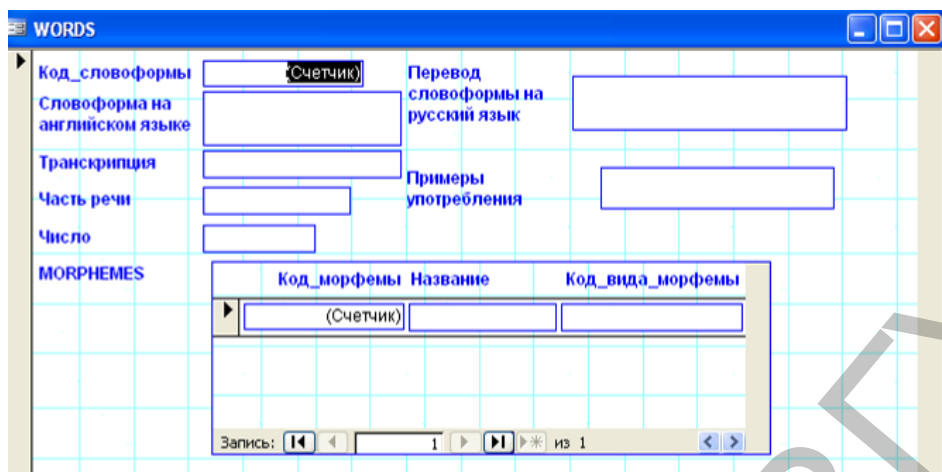


Рис. 7. Подчиненная форма «WORDS-MORPHEMES»

В подчиненную форму могут входить поля из нескольких таблиц: поля подчиненной таблицы, связанной с главной со стороны «один» или поля таблиц, связанных с подчиненной таблицей тоже со стороны «один». В форме значение ключа связи вводится и в главную таблицу, и в соответствующее поле подчиненной таблицы. В связанной форме подчиненная форма отображается на экране в виде кнопки. Если щелкнуть по этой кнопке мышью, то содержимое подчиненной формы будет показано на экране.

Задание 6. Создайте составную форму “WORDS- MORPHEMES”.

Задание 7 В таблицы БД “Словарь морфем английского языка” с помощью форм и обычным способом введите следующие данные:

Таблица 20. WORDS

Код_словоформы	Словоформа на английском языке	Транскрипция	Часть речи	Число	Перевод словоформы на русский язык	Примеры употребления
1	<i>underestimates</i>		verb	-	недооценивает	Alberta Health Services underestimates our respect for veterans.
2	<i>underestimating</i>		noun	-	недооценка	"Your friend's mistake was underestimating you," reassures Teddy.
3	<i>pounds</i>		noun	plural	фунты	He's definitely a great athlete, and he has good size at 6'4", 220 pounds .

Таблица 21. SPECIES MORPHEMES

Код_вида_морфемы	Обозначение	Класс морфем	Перевод класса морфем на русский язык	Вид морфемы	Перевод вида морфемы на русский язык
1	R	<i>roots</i>	корни	<i>root</i>	корень
2	L	<i>affixes</i>	аффиксы	<i>lexical suffix</i>	лексический суффикс
3	Pr	<i>affixes</i>	аффиксы	<i>prefix</i>	префикс
4	Gr	<i>affixes</i>	аффиксы	<i>grammatical suffix</i>	флексия или грамматический суффикс

Таблица 22. DICTIONARY

Код_словоформы	Код_морфемы
1	1
1	2
1	3
1	4
2	1
2	2
2	11
2	7

Таблица 23. MORPHEMES

Код_морфемы	Название	Код_вида_морфемы
1	under-	3
2	-estim-	1
3	-ate-	2
4	-s	4
5	-es	4
6	-ed	4
7	-ing	4
8	-er	4
9	-est	4
10	-en	4
11	-at-	2

Создание запросов к базе данных «Словарь морфем английского языка»

Одной из важных функций баз данных являются поиск и обработка данных по запросу пользователя. С помощью запросов можно отыскивать и просматривать определенные записи, обновлять и модифицировать данные, осуществлять расчеты, использовать результаты запросов для создания новых таблиц, форм, отчетов.

В СУБД Access существуют: *запросы на выборку*; *запросы с параметрами*; *перекрестные запросы*; *запросы на изменение* (обновление, добавление и удаление записей, создание таблиц по результатам запросов); *запросы SQL* (запросы на объединение, к серверу, управляющие и подчиненные запросы). Наиболее распространенный тип запросов – это *запросы на выборку*, в которых в формализованном виде представлен критерий поиска данных, необходимых конечному пользователю. Поиск может осуществляться по одной или сразу по нескольким взаимосвязанным таблицам. Результат поиска представляется в виде таблицы, в которую включены интересующие пользователя поля. *Запросы с параметром* позволяют пользователю с клавиатуры вводить изменяемые критерии поиска, однако при этом сама структура запроса не меняется.

В СУБД Access существуют два способа создания запросов: с помощью мастера; в режиме конструктора.

Создание запросов на выборку.

Задание 8. Требуется составить список всех словоформ словаря с указанием морфем и их обозначений. Создайте простой запрос на выборку с помощью Конструктора. В запрос включите все таблицы.

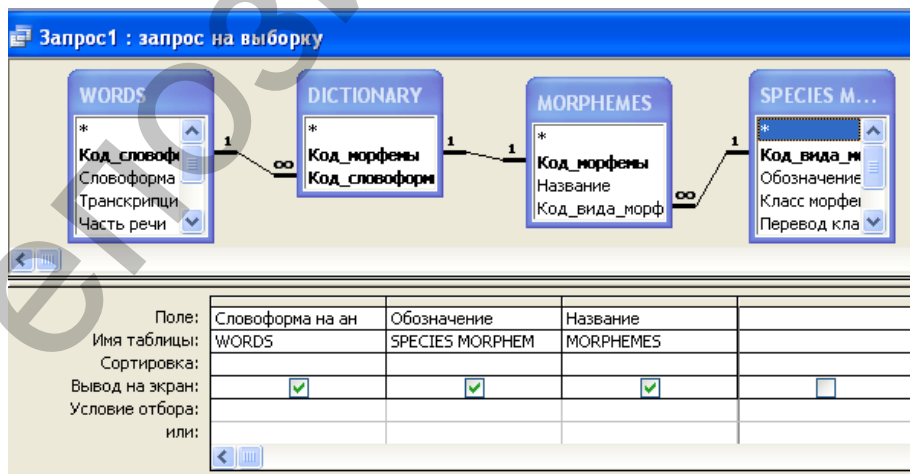


Рис. 8. Запрос на выборку в режиме Конструктора

Создание запросов с параметрами.

Запросы с параметрами целесообразно использовать тогда, когда по одному запросу необходимо периодически осуществлять рабо-

ту с данными при изменяющихся значениях в критерии поиска. При формировании запросов с параметрами для указания критерия отбора используются квадратные скобки.

Задание 9. Требуется найти словоформы, у которых есть флексия, причем она различна. В условии отбора запроса следует ввести выражение в квадратных скобках [Укажите флексию].

Создание отчетов в базе данных «Словарь морфем английского языка»

Использование отчетов является удобным и эффективным способом отображения результирующей информации. Отчеты можно создавать с помощью автоотчета, мастера отчетов, с помощью конструктора отчетов (отчет полностью формируется пользователем). В отчетах присутствует несколько разделов: заголовок, верхний и нижний колонтитулы, область данных и примечание отчета.

Задание 10 Необходимо подготовить отчет о тех словоформах, которые являются глаголами. Предварительно с помощью запроса (рис.9) необходимо отыскать нужную для отчета информацию, а затем создать отчет.

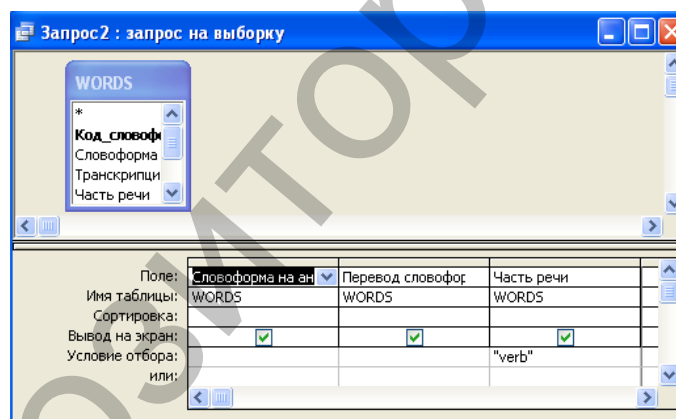


Рис. 9. Окно запроса о глаголах

По результатам запроса с помощью конструктора отчетов можно составить отчет, который приведен на рис. 10.

Английские глаголы

Словоформа на английском языке	underestimates
Перевод словоформы на русский язык	недооценивает

Рис. 10. Окно отчета о глаголах

Задание 11. Дополните таблицы словаря словоформами и их морфемами. Выполните запросы и проанализируйте результаты.

2. РЕКОМЕНДАЦИИ ПО САМОСТОЯТЕЛЬНОЙ РАБОТЕ

Самостоятельная работа является важнейшей составной частью учебного процесса и обязанностью каждого студента. Качество усвоения учебной дисциплины находится в прямой зависимости от способности студента самостоятельно учиться.

Содержанием самостоятельной работы студентов являются следующие ее виды:

- изучение понятийного аппарата дисциплины;
- индивидуальное изучение литературы по теме и подготовка устных сообщений и рефератов;
- выполнение дифференцированных и индивидуальных практических заданий с последующим сохранением и прикреплением результатов в Moodle;
- подготовка к контрольным тестам и опросам по основным разделам дисциплины;
- изучение тем, предназначенных для самостоятельной подготовки в соответствии с учебно-тематическим планом;
- самостоятельная работа студента при подготовке к зачету;
- самостоятельная работа студента в библиотеке.

Изучение дисциплины «ИТ в филологии» рассчитано на 72 учебных часа, из них – 36 часов аудиторных, самостоятельная работа – 36 часов. Итоговая форма контроля знаний – зачет.

Самостоятельная работа над проектами «Создание БД «Частотно-алфавитный словарь словоформ текста на иностранном языке»», «Создание БД «Словарь морфем немецкого (русского, французского, испанского) языка»», «Создание БД «Словарь цитат»», «Создание БД «Лингвистический терминологический словарь».

Тема работы: Проектирование базы данных по выбранному проекту.

Цель работы: создать таблицы базы данных, разработать схему данных, ввести данные в таблицы с помощью форм, выполнить запросы и подготовить отчеты средствами Microsoft Access..

Ход работы

1. Выбрать тему проекта.
2. Создать базы данных по выбранному проекту и придумать для них формы, запросы и отчеты.
3. Отправить на проверку преподавателю.
4. Ответить на контрольные вопросы (устно на занятии).

3. ТЕМАТИКА РЕФЕРАТОВ

Требования к рефератам **по предъявлению:**

Работа перед отправкой архивируется и именуется фамилией и аббревиатурой факультета, специальности и номером группы (латинскими символами!) автора. Например, IvanovER_FLF_RGF-11.

- 1) Технические, программные, алгоритмические и лингвистические средства информационных технологий.
- 2) Обзор сетевых ресурсов по корпусной лингвистике.
- 3) Специальные возможности программы MS Word для лингвистов.
- 4) Сравнение программ переводческой памяти (TRADOS, Déjà vu и т.п.).
- 5) Филологическая информатика как наука и учебная дисциплина.
- 6) Информационные ресурсы общества.
- 7) Лингвистические информационные ресурсы.
- 8) Эволюция операционных систем компьютеров различных типов.
- 9) Экспертные системы.
- 10) АРМ «Лингвист».

4. ПРИМЕРНЫЕ ВОПРОСЫ К ЗАЧЕТУ

1. Программы-браузеры. Определение URL.
2. Формальная и смысловая релевантность поиска.
3. Виды поиска: синтаксический и семантический.
4. Единица корпуса. Принципы отбора текстов для корпуса.
5. Классификация корпусов: «исследовательский корпус», «статический корпус», «параллельный корпус».
6. Структура машинной словарной статьи.
7. Определение базы данных. Основные способы организации баз данных.
8. Особенности электронных переводческих словарей АBBYY LINGVO и Multitran, их отличия от онлайн-переводчиков (Google, Yandex и т.п.).
9. Требования к специальным словарям.
10. Определение терминологического словаря. Отличия дескриптивных и нормативных терминологических словарей.
11. Этапы развития машинного перевода. Предредактирование и постредактирование.
12. Виды веб-ресурсов: образовательные порталы, электронные библиотеки, журналы в электронной версии, веб-квесты.
13. Проектирование базы данных. Основные понятия инфологической модели: сущность, связь и атрибут.
14. Отображение обязательности и необязательности связи на диаграмме инфологической модели.
15. Определение Схемы данных.
16. Способы ввода данных в таблицы БД.
17. Формы и отчеты в Microsoft Access.
18. Виды запросов, их характеристика.

ЛИТЕРАТУРА

1. Зубов, А.В. Информационные технологии в лингвистике : учеб. пособие для студ. лингв. фак-тов высш. учеб. заведений / А.В. Зубов, И.И. Зубова – М.: Издательский центр «Академия», – 2004. – 208 с.
2. Ривлина, А.А. Теоретическая грамматика английского языка : учебно-методическое пособие / А.А. Ривлина. – Благовещенск : Изд-во БГПУ, 2009. – 251 с.
3. Степанов, А.Н. Информатика: Учебник для вузов. 5-е изд. / А.Н. Степанов – Спб.: Питер, 2007. – 684 с.
4. Щипицина, Л.Ю. Информационные технологии в лингвистике : учеб. пособие / Л.Ю. Щипицина. – М. : ФЛИНТА : Наука, 2013. – 128 с.
5. Хроленко, А.Т. Современные информационные технологии для гуманитария : практ. руководство для студ., аспирантов, преподавателей-филологов / Л.Ю. Щипицина. – М. : ФЛИНТА : Наука, 2007. – 128 с.

Учебное издание

ОГАНДЖАНЫН Ольга Петровна

**МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ
ПО ИЗУЧЕНИЮ ДИСЦИПЛИНЫ
«ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ФИЛОЛОГИИ»**

Технический редактор *Г.В. Разбоева*

Компьютерный дизайн *И.В. Волкова*

Подписано в печать 2014. Формат 60x84¹/₁₆. Бумага офсетная.

Усл. печ. л. 2,90. Уч.-изд. л. 2,16. Тираж экз. Заказ .

Издатель и полиграфическое исполнение – учреждение образования
«Витебский государственный университет имени П.М. Машерова».

Свидетельство о государственной регистрации в качестве издателя,
изготовителя, распространителя печатных изданий

№ 1/255 от 31.03.2014 г.

Отпечатано на ризографе учреждения образования
«Витебский государственный университет имени П.М. Машерова».

210038, г. Витебск, Московский проспект, 33.