Л.Ю. Щипицина

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ЛИНГВИСТИКЕ

Учебное пособие

Москва
Издательство «ФЛИНТА»
Издательство «Наука»
2013

УДК 800(075.8)
ББК 81.1я73
Щ85

Р е ц е н з е н т ы:

д-р филол. наук, проф., зав. кафедрой теории перевода
и межкультурной коммуникации Воронежского
государственного университета *В.Б. Кашкин*;

канд. техн. наук, доцент кафедры информационных технологий
и автоматизированных систем Московского государственного института
электроники и математики *Э.С. Клышинский*

Щипицина Л.Ю.

Щ85 Информационные технологии в лингвистике : учеб. пособие / Л.Ю. Щипицина. — М. : ФЛИНТА : Наука, 2013. — 128 с.

ISBN 978-5-9765-1431-7 (ФЛИНТА)
ISBN 978-5-02-037776-9 (Наука)

В учебном пособии излагаются основы курса «Информационные технологии в лингвистике», приводятся задания для организации самостоятельной работы студентов и глоссарий, включающий необходимые понятия курса.

Для преподавателей и студентов филологических и лингвистических специальностей.

УДК 800(075.8)
ББК 81.1я73

ISBN 978-5-9765-1431-7 (ФЛИНТА)
ISBN 978-5-02-037776-9 (Наука)

Содержание

Предисловие

Информационные технологии в настоящее время являются неотъемлемой частью любой сферы профессиональной деятельности, в том числе лингвистики. И если когда-то использование компьютеров и соответствующих программ в лингвистических исследованиях, переводе и в обучении языку не являлось обязательным, то сегодня уже со студенческой скамьи будущим преподавателям иностранных языков, переводчикам и лингвистам-исследователям необходимы компетенции, связанные с использованием информационных технологий в своей профессиональной сфере деятельности.

Первичному знакомству с возможностями информационных технологий в лингвистике служит настоящее учебное пособие, которое предназначено для студентов лингвистических специальностей бакалавриата младших курсов. Пособие соответствует рабочей программе дисциплины «Информационные технологии в лингвистике» и может быть использовано в качестве основного источника литературы по этой дисциплине.

Именно полный охват тем курса, подлежащих изучению студентами, а также наличие системы заданий и упражнений, облегчающих формирование у обучающихся необходимых компетенций, отличает данное пособие от других подобных изданий.

Пособие включает три основных части, библиографический список и приложения.

В основное содержание пособия входит часть 1 «Основные понятия», часть 2 «Области применения информационных технологий в лингвистике» и часть 3 «Прикладные разделы компьютерной лингвистики». Каждая часть содержит несколько разделов, включающих перечень основных теоретических вопросов, рассматриваемых в разделе, их краткое изложение, вопросы для обсуждения на семинарских занятиях, список литературы для самостоятельной подготовки студентов по теме раздела, упражнения и лабораторные работы. Для выполнения лабораторных работ требуются главным

образом базовые программы операционной системы и ресурсы Интернета, что до минимума сводит необходимость привлечения дополнительного программного обеспечения в ходе изучения курса.

В библиографическом списке приводится литература, использованная при подготовке пособия, а также список интернет-ресурсов, который может быть дополнен студентами при работе над курсом.

В приложении приводится глоссарий с определениями необходимых теоретических понятий курса, сформулированных автором пособия с опорой на различные источники, список тем, предлагаемых студентам для более глубокой проработки в виде индивидуальных докладов, а также тест для проверки знаний по курсу, снабженный ключами, что позволяет использовать тест для индивидуальной работы студентов.

Содержание и учебно-методический аппарат пособия позволяют рассматривать его как базовое в освоении возможностей информационных технологий в лингвистике. В дальнейшем предусматривается углубленное изучение отдельных разделов курса («Машинный перевод», «Автоматический анализ текста», «Компьютерная лингводидактика» и т.п.) в зависимости от профиля подготовки обучающегося в рамках специальных дисциплин профессионального цикла бакалавриата и магистратуры.

Часть 1
ОСНОВНЫЕ ПОНЯТИЯ

1.1. Лингвистика. Язык

Лингвистика как наука о закономерностях строения и развития естественного языка. Понятие теоретической и прикладной лингвистики. Соотношение прикладной и компьютерной лингвистики.

Язык как знаковая система. Понятие естественного и искусственного языка. Виды искусственных языков.

Изучение возможностей применения информационных технологий в лингвистике предполагает знание основных понятий соответствующей области знания, среди которых можно выделить понятия из сферы лингвистики (язык, лингвистика, компьютерная лингвистика и т.п.) и информатики (информация, алгоритм, модель и др.). Знакомство с этими понятиями начнем с лингвистических терминов, характеризующих непосредственную профессиональную область деятельности лингвистов, преподавателей иностранных языков и переводчиков.

Лингвистика (или языкознание) традиционно понимается как наука о естественном человеческом языке [9, 28]. Лингвистов занимают вопросы строения языка (выделение в нем фонетического, лексического, грамматического уровня и уровня текста), социального варьирования языка, вопросы порождения и понимания языковых высказываний, принципы функционирования языка в обществах разных типов, происхождения и развития языка и другие его аспекты [13, 618—622].

В зависимости от изучаемого аспекта языка, национальной традиции и научной методологии выделяются различные разделы лингвистики, например структурная лингвистика, социолингвистика, психолингвистика и т.п.

Чтобы определить раздел лингвистики, наиболее тесно связанный с использованием информационных технологий, целесообразно

обратиться к разграничению теоретической и прикладной лингвистики.

Теоретическая (или фундаментальная) лингвистика — это область языкознания, направленная на объективное установление состояния отдельного языка, его истории и закономерностей. Эта область лингвистики призвана ответить на вопрос «Каков язык?» [37, 214—215].

Прикладная лингвистика развивается с конца 20-х годов XX в. и является областью языкознания, связанной с разработкой методов решения практических задач использования языка [13, 397]. Прикладная лингвистика отвечает на вопрос «Как лучше использовать язык?».

Следует отметить, что в России и за рубежом сложились разные интерпретации понятия прикладной лингвистики. Если за рубежом в 1930—1940-е годы под прикладной лингвистикой (*Applied Linguistics*) прежде всего понимается процесс обучения иностранному языку, методика его преподавания, особенности описания грамматики для учебных целей, то в России начиная с 1950-х годов, прикладная лингвистика ассоциируется с компьютерными технологиями и автоматическими системами обработки информации [4, 6]. В связи с этим в русскоязычной научной традиции прикладная лингвистика нередко рассматривается как синоним компьютерной / вычислительной / автоматической / инженерной лингвистики.

На современном этапе развития науки в рамках прикладной лингвистики выделяется несколько направлений по оптимизации использования языка, которые объединяются исследователями в две большие группы: традиционные («вечные») и новые.

К традиционным направлениям и соответствующим задачам прикладной лингвистики относятся:

- создание и совершенствование письменностей;
- создание систем транскрипции устной речи;
- создание систем транслитерации иноязычных слов;
- создание систем стенографии;
- создание систем письма для слепых;

- упорядочение, унификация и стандартизация научно-технической терминологии;
- изучение процессов и создание правил образования названий новых изделий, товаров, химических веществ;
- разработка методов адекватного преобразования текстов в иноязычную форму (перевода);
- совершенствование методики преподавания языков и др. [13, 397].

Новыми задачами прикладной лингвистики считаются:

- разработка лингвистических основ машинного перевода;
- автоматическое индексирование и аннотирование документов;
- автоматический анализ текстов;
- автоматический синтез текстов;
- создание словарей-тезаурусов для автоматического поиска информации и др. [ср. 36].

Именно те области прикладной лингвистики, которые связаны с привлечением компьютеров для решения практических задач использования языка, являются предметом компьютерной лингвистики, оформившейся в 1960-е годы как особое научное направление.

Компьютерную лингвистику можно определить как область использования компьютерных инструментов — программ, технологий организации и обработки данных — для моделирования функционирования языка в тех или иных условиях, а также сферу применения компьютерных моделей языка в лингвистике и смежных с ней дисциплинах [4, 13].

В связи с тем, что язык представляет собой весьма сложное образование, в компьютерной лингвистике сложились и развиваются различные направления, примерно сопоставимые с отдельными уровнями языка, с процессами порождения и восприятия языковых сообщений или другими видами человеческой деятельности, связанной с языком. Соответственно, к направлениям компьютерной лингвистики относятся:

- автоматический анализ текстов;
- автоматический синтез текстов;

- создание и поддержка автоматических словарей;
- создание автоматизированных информационно-поисковых систем;
- машинный перевод;
- создание автоматических систем обучения языку;
- автоматическая атрибуция и дешифровка анонимных текстов;
- создание лингвистических баз данных;
- разработка программных инструментов для решения задач теоретической и прикладной лингвистики и т.д. [20; 53 и др.].

Лингвистика в целом и компьютерная лингвистика в частности имеют дело с языками различного типа и их отдельными уровнями. Язык в наиболее общем виде определяется как знаковая система, используемая для общения в некотором социуме [13, 604; 29, 5].

Различают естественные и искусственные языки. Естественный язык — это исторически сложившаяся и используемая в определенной этнической группе или национальном государстве знаковая система. Примерами естественных языков выступают русский и английский (принадлежащие к индоевропейской языковой семье) или финский и эстонский (принадлежащие к финно-угорской языковой семье).

Искусственные языки представляют собой знаковые системы, искусственно создаваемые в тех областях, где применение естественных языков менее эффективно или невозможно. Среди искусственных выделяются неспециализированные (или международные) языки (эсперанто, волапюк и др.) и специализированные языки. К последним относятся языки науки (математики, логики, химии и т.д., создание которых началось в XVI в.) и языки человеко-машинного общения (получающие распространение в специальных областях человеческой деятельности, связанной с облегчением диалога человека и компьютера, начиная с 1940-х годов) [13, 201—202].

Примеры языков человеко-машинного общения простираются от простейших систем символического кодирования (ассемблеров) до специализированных языков программирования (C++, Java, Python, ErLang и др.). К 1980-м годам в мире насчитывалось около 500 языков программирования [13, 202]. В настоящее время активно

используется примерно столько же, хотя общее количество известных языков программирования достигает нескольких тысяч [55]. Эти факты свидетельствуют об остроте проблемы человеко-машинного общения и о множестве подходов к ее решению.

Подводя итог разделу, констатируем, что лингвистикой следует считать науку о закономерностях происхождения, строения и функционирования естественного человеческого языка. Предметом лингвистики и компьютерной лингвистики как ее особого раздела выступает язык — знаковая система, используемая с различными целями.

Вопросы для обсуждения

1. Что такое лингвистика? Назовите ее разделы. В каком разделе лингвистика имеет дело с информационными технологиями?
2. Можно ли считать синонимами прикладную и компьютерную лингвистику? Аргументируйте свой ответ.
3. Перечислите основные направления компьютерной лингвистики. Расскажите об одном из направлений.
4. Сравните разные определения языка. Выделите в них ключевые слова. Составьте на основе повторяющихся ключевых слов свое определение языка.
5. Подумайте, с естественным или искусственным языком имеет дело компьютерная лингвистика?
6. Какие виды естественных и искусственных языков вам известны? Приведите примеры естественных и искусственных языков разных видов.

Рекомендуемая литература

1. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. 3-е изд. М.: ЛКИ, 2007. С. 6—8, 20.
2. Беляева Л.Н. Лингвистические автоматы в современных гуманитарных технологиях: учеб. пособие. СПб.: Книжный Дом, 2007. С. 36—40.
3. Большой энциклопедический словарь. Языкознание. М.: Большая Российская энциклопедия, 1998. С. 201—202; 604—606, 618—622.
4. Всеволодова А.В. Компьютерная обработка лингвистических данных: учеб. пособие. 2-е изд., испр. М.: Флинта: Наука, 2007. С. 63—64.
5. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 5—7.

Упражнения

1. Определите статистические показатели приведенного ниже текста смешанного языкового типа.

Проекты Cíbola/Oleada реализуют обширные компьютерные системы лингвистического анализа текстов, представленных в Unicode. Компоненты системы включают средства работы с мультиязыковыми текстами (MUTT), построения конкорданса (XConcord) для текстов на более чем 16 языках, статистического анализа, автоматического перевода, различные словари и тезаурусы. Некоторые версии этих компонентов доступны для бесплатной загрузки после процедуры формальной регистрации. Все компоненты реализованы в среде X11 Window System для SunOs и Solaris (*источник: Проекты Cíbola/Oleada http://rvb.ru/soft/catalogue/c01.html*).

Слов	
Символов (без пробелов)	
Символов (с пробелами)	
Символов в латинской графике	
Чисел	
Средняя длина слов	

2. Какому языку соответствует средняя длина слов текста смешанного типа, приведенного в задании 1? Для выполнения задания вычислите среднюю длину слов русского языка из приведенного текста и среднюю длину слов в латинской графике.
3. Определите, каким языкам соответствуют следующие специфические буквы, буквосочетания и слова:
   а) Ø ö ß ρ ω ë š,
   б) th sch šč,
   в) et, the, der, och, için.
4. Создайте диагностический словарь для определения языка на материале текстов на двух разных языках (на ваш выбор). Для этого заполните следующую таблицу.

Таблица

Критерий	Язык 1:	Язык 2:
Типичные артикли		
Указательные местоимения		

Окончание таблицы

Критерий	Язык 1:	Язык 2:
Местоимения 3-го лица		
Отдельные формы вспомогательных глаголов		
Основные предлоги и союзы		
Другие частотные слова		

5. Дополните таблицу встречаемости букв в распространенных европейских языках [Всеволодова 2007: 64], добавив в нее данные по русскому языку. Используйте для этого любой текст на русском языке объемом не менее 100 символов.

6. Прочитайте несколько фраз на эсперанто. Назовите морфологические диагностические показатели этого языка, учитывая, что существительные и прилагательные на эсперанто всегда имеют одни и те же окончания.

Рус.	*Эсперанто*
зеленое дерево	verda arbo
старый человек	maljuna viro
хороший друг	bela amiko

1.2. Информация. Информационные технологии

Информация как предмет изучения информатики и кибернетики. Понятие информационных технологий в лингвистике.

Виды информации. Способы кодирования и носители информации. Информационные революции.

Понятие модели и алгоритма в информатике. Понятие искусственного интеллекта.

Одним из основных назначений языка является его использование для передачи информации между людьми. Поэтому, говоря о языке, невозможно обойти вниманием и понятие информации.

Информация в обыденном понимании трактуется как сведения о положении дел в окружающем мире, его свойствах, протекающих в нем процессах и т.п. [31]. В специальных науках, изучающих информацию, это понятие определяется несколько иначе: как последовательность сигналов или символов некоторого алфавита, кодирую-

щая некоторое сообщение без учета смыслового содержания этого сообщения (в теории передачи информации) или как содержание, которое получено из внешнего мира и позволяет адекватно реагировать живому организму (или технической системе) на окружающую среду (в кибернетике) [16, 11—12].

Обобщая различные определения информации, можно предложить следующую трактовку этого понятия: *информация* — это сведения об окружающем мире, передаваемые человеком, живыми организмами или техническими системами для регулирования своего поведения в окружающей среде.

Роль информации в современном обществе исключительно велика. Информация, кодируемая с помощью языка, превращается в знания; знания же передаются от поколения к поколению, тем самым обеспечивая преемственность общественных устоев.

Информация может кодироваться вербально или невербально. Различие способов кодирования информации (аудитивный, тактильный, визуальный, густический и т.д.) обусловливает множество способов ее представления:

- тексты;
- рисунки, чертежи, фотографии;
- световые или звуковые сигналы;
- электрические и нервные импульсы;
- жесты и мимика;
- запахи и вкусовые ощущения;
- хромосомы, посредством которых передаются по наследству признаки и свойства организмов, и т.д.

Способов представления информации, как показывают примеры, достаточно много. Но поскольку человек может воспринимать информацию лишь с помощью собственных органов чувств, целесообразно классифицировать виды информации именно на этом основании. По тому, какими органами чувств воспринимаются и какой сигнальной системой закодированы сведения об окружающем мире, можно выделить звуковую, вкусовую, тактильную, визуально-образную и визуально-символическую информацию. Именно последние

два вида информации являются наиболее значимыми для современного человека, при этом если в XX в. человек имел дело в основном с визуально-образной, то в XXI в. наиболее значимой становится визуально-символическая информация.

Символ (греч. σύμβολον) — это знак, обозначающий некоторый предмет или явление. В лингвистике символами считаются в первую очередь слова, поскольку именно слово является минимальной единицей, способной обозначать предметы и явления окружающего мира. В информатике символами считаются главным образом буквы, знаки препинания, цифры и другие знаки печатного текста, а также звуковые знаки — фонемы — устного текста, являющиеся составляющими алфавитов и фонетических систем различных естественных и искусственных языков. Эти символы складываются в слова и предложения, кодирующие передаваемую информацию.

Процессы, связанные с определенными операциями над информацией, называются информационными процессами. В настоящее время над информацией можно производить следующие операции:

создавать	принимать	комбинировать
хранить	передавать	копировать
искать	воспринимать	формализовать
измерять	использовать	делить на части
упрощать	разрушать	обрабатывать
собирать	распространять	преобразовывать

В связи с постоянным увеличением количества используемой людьми информации на определенном этапе развития общества потребовалось привлечение специальных технических средств для ее обработки и хранения. Принципиальные изменения в способах фиксации и передачи информации, связанные с изобретением новых технических средств получили название информационных революций. Исследователями выделяются три информационные революции [28, 404—405]:

1) ок. 3000 лет до н.э. — изобретение письменности (шумерская клинопись): информация может накапливаться;
2) 1453 г. — изобретение книгопечатания: информация становится массово доступной;

3) начало 1970-х годов — создание персональных ЭВМ и телекоммуникационных сетей: информация может автоматически обрабатываться и доставляться в электронном виде с высокой скоростью.

Третья информационная революция в значительной степени стимулировалась тем, что в середине XX в. появляются специальные науки, изучающие информацию: информатика и кибернетика. *Информатика* — это наука о накоплении, обработке и передаче информации с помощью ЭВМ. Наука об управлении, связи и переработке информации называется кибернетикой.

Именно в рамках теории информации (математической теории связи) для иллюстрации информационного обмена, осуществляемого с помощью технических средств, К. Шенноном и У. Уивером была предложена наглядная модель (рис. 1).

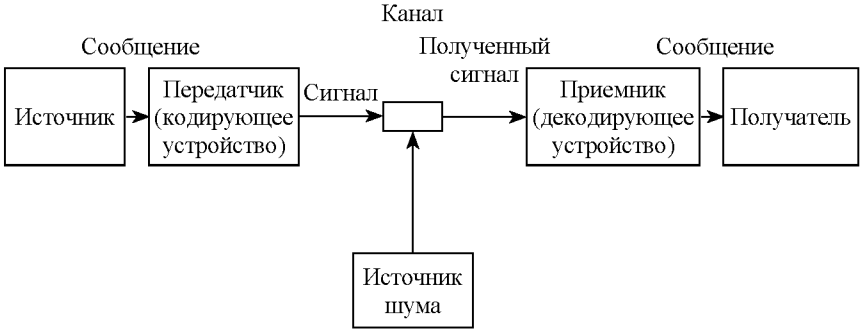


Рис. 1. **Модель К. Шеннона и У. Уивера [33, 131]**

Особо значимым для информационных технологий представляется указание в данной модели на кодирующее и декодирующее устройство, поскольку одной из важных задач информатики является перевод информации, закодированной в «человеческих» символах, в информацию, понятную компьютерам, и наоборот.

Компьютеры в информационном обмене становятся средством кодирования, обработки, хранения и передачи больших массивов символьной информации. Совокупность законов, методов и средств получения, хранения, передачи, распространения и преобразования информации с помощью компьютеров получило обозначение «*информационные технологии*».

При сужении этого понятия для его использования в особой профессиональной сфере (лингвистика) получаем сочетание «*информационные технологии в лингвистике*», понимаемое как совокупность законов, методов и средств получения, хранения, передачи, распространения и преобразования информации о языке и законах его функционирования с помощью компьютеров [20, 8].

Одной из задач соответствующей области знания является сравнение способов кодирования информации человеком и компьютером.

Под *кодированием* в целом понимается процесс представления информации в виде последовательности условных обозначений. Иными словами, кодирование — это сопоставление объектов и отношений между ними с символами или словами какого-либо языка [16, 39—40].

В процессе кодирования соотношение слова (символа) и его значения обычно называется *семантикой*, правила, выражающие общие синтаксические свойства слов и групп слов, позволяющие производить и/или описывать правильные предложения языка — *грамматикой* [11, 98; 51, 19].

О способах кодирования информации человеком говорилось выше. Компьютер может обрабатывать все известные виды информации, включая:

- числовую,
- буквенную (вербальную),
- графическую,
- звуковую,
- видеоинформацию.

Информация в компьютере представлена в двоичном коде, алфавит которого состоит из двух цифр (0 и 1).

Так, *числовая информация* используемой человеком десятичной системы счисления предстает в ЭВМ в виде следующих сочетаний символов 0 и 1:

0 — 0	4 — 100	8 — 1000
1 — 1	5 — 101	9 — 1001
2 — 10	6 — 110	10 — 1010
3 — 11	7 — 111	

Для кодирования компьютером *вербальной информации* изначально использовался код ASCII (American Standard Code for Information Interchange). Для кодирования одного символа в этом коде требуется 1 байт (или 8 битов). В целом в ASCII можно закодировать 256 символов, при этом каждому символу ставится в соответствие уникальный десятичный код от 0 до 255. Так, запись слова «КОМПЬЮТЕР» в двоичном коде выглядит следующим образом (табл. 1).

Таблица 1

Двоичные коды символов, составляющих слово «КОМПЬЮТЕР» [43, 62]

1	2	3	4	5	6	7	8	9
К	О	М	П	Ь	Ю	Т	Е	Р
10001010	10001110	10001100	10001111	10011100	10001110	10010010	10000101	10010000

Для кодирования данного слова в памяти компьютера потребуется 9 восьмибитовых комбинаций цифр, т.е. 9 байтов. Следует помнить о том, что двоичные коды этого же слова, написанного строчными буквами, будут иными.

В настоящее время для увеличения количества символов, которые могут быть зашифрованы в одной и той же системе кодирования, используется стандарт UNICODE, в котором для кодирования одного символа используется два байта.

Для кодирования *графической информации* обычно используется 2 способа — представление рисунка в виде растрового или векторного изображения. Растровое изображение формируется из определенного количества строк, содержащих определенное количество точек (пикселей). Векторное изображение — графический объект, состоящий из элементарных графических объектов, например отрезков и дуг. Положение этих элементарных объектов определяется координатами точек и длиной радиуса.

Кодирование звуковой информации опирается на материальные характеристики этой информации. Известно, что звук представляет собой звуковую волну с непрерывно меняющейся амплитудой и частотой звучания. Чем больше амплитуда сигнала, тем он громче, чем больше частота сигнала, тем выше тон. Визуально представить зву-

ковую волну помогает фонограмма, т.е. зафиксированные специальными приборами и отражаемые, к примеру, на экране монитора колебания звуковой волны (рис. 2).

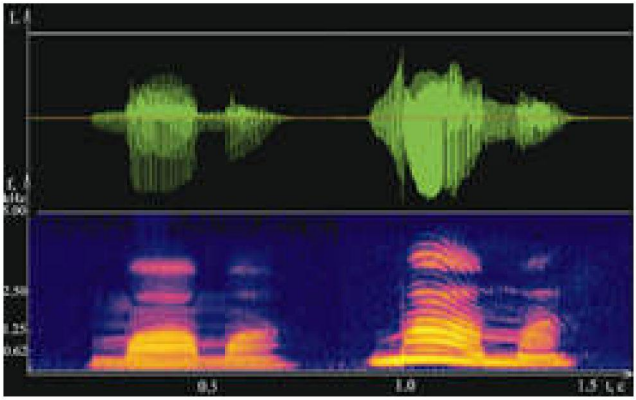


Рис. 2. **Визуальное представление слова «мама» [2]**

При *кодировании видео* к звуковой информации добавляются визуальные изображения, представляемые в виде множества отдельных кадров, плавно переходящих один в другой на временной оси.

Для компьютерной обработки лингвистических данных важно иметь представление о компьютерной лингвистической модели и об алгоритме решения лингвистических задач.

Моделью обычно считают материальный или идеальный образ некоторой совокупности предметов или явлений, заменяющий реальные предметы и явления и включающий только их наиболее существенные признаки [43, 38]. Примерами материальных моделей выступают рисунки или трехмерные изображения молекул в химии, солнечной системы в астрономии, организма человека в анатомии.

Лингвистические модели являются большей частью идеальными конструктами, позволяющими раскрыть особенности строения и функционирования языка, производство и восприятие речи и текста [20, 14]. Простейшие лингвистические модели иллюстрируют строение слова из фонем, предложения из именных и глагольных групп, текста из единиц сюжета. Так, базовыми элементами текста в сюжетной грамматике выступают экспозиция, событие и эпизод [4, 27]. Сложные лингвистические модели включают большее количество состав-

ляющих различных уровней и отличаются комплексными целями (ср. параграф 2.4 «Автоматический анализ и синтез текста»).

Построение компьютерных лингвистических моделей предполагает выполнение некоторой последовательности действий. Формализованное описание такой последовательности действий, приводящей к решению поставленной задачи, называется *алгоритмом* [43, 40]. Алгоритмы могут быть записаны в виде вербальных инструкций, блок-схем, таблиц или на языках программирования. Примеры алгоритмов различного рода см. в работе [20, 18—19, 36—37].

С 1970-х годов различные подходы к моделированию человеческой деятельности в различных сферах и предметных областях интегрируются в усилиях по созданию искусственного интеллекта. Под *искусственным интеллектом* (англ. *Artificial Intelligence*) понимается междисциплинарная область исследований, связанная с созданием сложных человеко-машинных и робототехнических систем [13, 14].

Подводя итог содержанию данного раздела, констатируем: информация, являющаяся непременным условием существования человеческого общества, представляет собой сведения об окружающем мире, передаваемые человеком, живыми организмами или техническими системами для адекватной реакции на изменения в окружающей среде. Компьютерные инструменты получения, хранения, передачи, распространения и преобразования информации, а также соответствующие законы и методы получили обозначение информационных технологий. Если с помощью компьютеров мы получаем, храним, передаем и распространяем любую информацию, касающуюся языка и законов его функционирования, мы имеем дело с информационными технологиями в лингвистике.

Вопросы для обсуждения

1. Сопоставьте разные определения информации. Какое из определений, на ваш взгляд, лучше всего подходит к лингвистике?
2. Сравните свойства информации, выделяемые в разных источниках.
3. Как соотносятся информация, сообщение и данные?
4. Назовите основные этапы развития информационных технологий.

5. В чем ученые видят будущее информационных технологий? Что вы думаете по этому поводу?
6. Что такое задача и правило? Как эти понятия связаны с алгоритмом?
7. Каковы свойства алгоритмов?

Рекомендуемая литература

1. Всеволодова А.В. Компьютерная обработка лингвистических данных: учеб. пособие. 2-е изд., испр. М.: Флинта: Наука, 2007. С. 9—16.
2. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 7—19.
3. Степанов А.Н. Информатика: учеб. пособие. СПб.: Питер, 2006. С. 35—42.

Упражнения

1. Найдите лишнее в приведенном ниже списке. Решите данную задачу с точки зрения компьютерной семантики и компьютерной грамматики.

   Ландыш, левкой, лаватера, лютик, люпин, ромашка, липа.

2. Дайте определения элементам следующих синтаксических моделей, примеры цит. по: [Апресян 1966: 167—168].

   a) $A_nN_n \leftrightarrow N(A)_nN_g$ *(быстрое движение ↔ быстрота движения)*

   b) $VN_a \leftrightarrow N(V)_nN_g$ *(прибавляю число ↔ прибавление числа)*

   c) $N^1{}_nN^2{}_g \leftrightarrow A(N^2)_nN^1{}_n$ *(права автора ↔ авторские права)*

   d) $V_{в}N_a \leftrightarrow N(V)_{n}{}_{в}N_a$ *(возвожу в степень ↔ возведение в степень)*

   e) $DV \leftrightarrow A(D)_nN(V)_n$ *(сильно желать ↔ сильное желание)*

   $A_n = \ldots\ N_n = \ldots\ N_g = \ldots\ N(A)_n = \ldots\ N(V)_n = \ldots\ V = \ldots\ D = \ldots$

3. По образцу задания 2 составьте модели следующих трансформаций: *визит врача ↔ врачебный визит, ароматный сад ↔ аромат сада, выхожу из дома ↔ выход из дома.*

4. Приведите примеры словосочетаний (а) и предложения (б) на русском языке, описываемых следующими моделями:

   а) $N^1{}_n$ из $N^2{}_g \leftrightarrow A(N^2)_nN^1{}_n$

   б) $A_nN_nVDA_aN_a$

1.3. Аппаратное и программное обеспечение информационных технологий в лингвистике

Компьютер и периферийные устройства как аппаратная основа информационных технологий. Системное и прикладное программное обеспечение. Лингвистические ресурсы (*lingware*). Автоматизированное рабочее место лингвиста.

Для выполнения объемных расчетов над лингвистическими данными, а также для лингвистического моделирования удобно использовать электронные вычислительные машины (или компьютеры). *Компьютер* — это электронное устройство, служащее для автоматического создания, обработки, передачи и воспроизводства информации по созданным человеком алгоритмам (программам), написанным на понятном для машины языке [43, 42; 15, 22].

Как следует из приведенного определения, в использовании компьютеров сочетается аппаратное (*hardware*) и программное обеспечение (*software*) информационных технологий.

К аппаратному обеспечению относится сам компьютер (стационарный или переносной), а также периферийные устройства, служащие для ввода/вывода информации в компьютер пользователем (клавиатура, мышь, монитор, принтер и т.д.) или для соединения компьютера с другими устройствами (например, модем).

Программное обеспечение — это компьютерные программы, представляющие собой последовательность написанных на машинном языке команд, служащие для управления аппаратными средствами или для выполнения различных операций над информацией, и соответствующая документация.

В зависимости от назначения программных средств различают системное и прикладное программное обеспечение. *Системные программы* служат управлению работой аппаратных средств и включают операционные системы, утилиты, драйверы и некоторые другие виды программ. *Прикладные программы* предназначены для конечного пользователя и позволяют ему выполнять различные операции над информацией: создавать и обрабатывать текст (текстовые редакторы), обрабатывать графические изображения (графические редакторы), работать над звуковой и видеоинформацией (мультимедий-

ные программы), создавать электронные таблицы для обработки статистических данных (электронные таблицы) и т.д. Для лингвиста особенно полезными являются такие виды прикладных программ, как электронные переводчики и словари, а также мультимедийные обучающие программы.

Наряду с аппаратным и программным обеспечением (ПО) информационных технологий некоторые исследователи используют также понятие lingware (или *linguware*), которым обозначаются все лингвистические компьютерные ресурсы (грамматические справочники, словари, энциклопедии, лингвистические базы данных и т.п.) [ср. 8, 27, 31; 59].

Совокупность аппаратных, программных и лингвистических средств, необходимых для автоматической обработки лингвистических данных, обозначим понятием *автоматическое рабочее место* (АРМ) *лингвиста* [22, 258]. АРМ лингвиста будет включать сам компьютер, операционное и базовое прикладное ПО, а также всевозможные лингвистические компьютерные ресурсы, касающиеся родного и изучаемых иностранных языков.

В зависимости от специализации АРМ лингвиста может дополняться прикладными программами и лингвистическими ресурсами, связанными с переводом или обучением иностранному языку. Задачей обучающихся является постоянная актуализация своего АРМ, включающая поддержание современного состояния аппаратного и программного обеспечения, а также постоянное пополнение собственной лингвистической ресурсной базы, т.е. поиск, сохранение, приобретение или создание лингвистических справочников, словарей и баз данных.

Вопросы для обсуждения

1. Опишите строение компьютера и охарактеризуйте периферийные устройства.
2. Дайте определение системному и прикладному программному обеспечению. Определите понятия операционной системы, утилиты и драйвера.
3. Приведите классификацию прикладных компьютерных программ. Дайте их краткую характеристику и приведите примеры основных видов прикладных компьютерных программ.
4. Охарактеризуйте текстовый процессор и его лингвистические функции.

5. Охарактеризуйте специальные компьютерные программы, разработанные для лингвистических целей.
6. Опишите лингвистические ресурсы компьютерной лингвистики (*lingware*).

Рекомендуемая литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. С. 97—99.
2. Всеволодова А.В. Компьютерная обработка лингвистических данных: учеб. пособие. 2-е изд., испр. М.: Флинта: Наука, 2007. С. 22—26, 37—53.
3. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие для студ. вузов. М.: Академия, 2004. С. 19—22.
4. Овчинникова И.Г., Угланова И.А. Компьютерное моделирование вербальной коммуникации: учебно-метод. пособие. М.: Флинта: Наука, 2009. С. 92—102.
5. Степанов А.Н. Информатика: учеб. пособие. СПб.: Питер, 2006. С. 42—43, 80—85, 106—111.
6. Чухарев Е.М. Компьютерные технологии в лингвистических исследованиях: указания по выполнению домашнего задания. Архангельск, 2009. С. 2—4.

Упражнение

Определите, к какому виду прикладных программ относятся перечисленные ниже программные продукты.

1) Текстовые редакторы
2) Графические редакторы
3) Электронные таблицы
4) Веб-редакторы
5) Веб-браузеры

Opera, MS Excel, MS FrontPage, Adobe Photoshop, Corel WordPerfect

Лабораторная работа 1

Простой поиск

Найдите в Интернете текст *Alice's Adventures in Wonderland by Lewis Carroll* (например, на сайте www.gutenberg.org/ebooks/11). Сохраните его на свой

компьютер в формате MS Word. Выполните задания на простой поиск в этом документе и внесите результаты поиска в таблицу.

Задание	Ответ
1. Сколько раз в тексте встречается слово *child* (в разных формах)?	
2. Сколько раз в тексте встречается слово *child* именно в этой форме?	
3. Приведите один из контекстов использования в тексте слова *beautiful*	
4. В какой орфографии (британской или американской) представлен текст?	Ответ: Проверочное слово:

Лабораторная работа 2

Поиск с подстановочными знаками

Выполните поиск с подстановочными знаками по тексту *Alice's Adventures in Wonderland*. Внесите результаты поиска в таблицу.

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

Лабораторная работа 3

Форматирование документа и проверка правописания в MS Word 2007

Для форматирования возьмите текст вашего доклада или подготовленного к семинару выступления.

1) В разделе «Главная» выберите «Выделить все». Текст должен оставаться выделенным во время дальнейших действий 2—5 (для этого не следует нажимать кнопки мыши в пространстве текста, а работать только с пунктами верхнего меню).

2) В разделе «Разметка страницы» в меню «Параметры страницы» назначьте размеры полей: верхнее — 2, левое — 3, нижнее — 2, правое — 2. Выберите «Ориентация страницы — книжная». Нажмите кнопку «ОК».

3) В этом же разделе «Разметка страницы» выберите пункт «Расстановка переносов» и нажмите «Авто».

4) Перейдите в раздел «Главная». Назначьте шрифт Century Schoolbook. Размер шрифта 12.

5) В разделе «Главная» выберите вкладку «Абзац». Поставьте выравнивание документа по ширине. Первая строка — отступ 0,6. Интервал «Перед и после» — 0, «междустрочный» — 1,5.

6) Щелкните кнопкой мыши в пространстве основного текста (выделение текста снимется).

7) Оформите титульную страницу документа, которая должна включать: название вуза и института, ФИО и № группы докладчика, дату устного выступления, тему, название курса и ФИО преподавателя, город и год.

   После оформления титульной страницы нажмите раздел «Вставка — Разрыв страницы» (или нажмите одновременно клавиши Ctrl + Enter) (основной текст доклада будет начинаться со 2-й страницы документа). Равномерно распределите информацию на титульной странице.

7) В разделе «Вставка» нажмите «Номер страницы». Выберите номер «Верху страницы — простой номер 3 (справа). Поставьте галочку в строке «Особый колонтитул для первой страницы (номер на ней не будет отображаться, но в нумерацию будет включен). Проверьте, чтобы нумерация страниц начиналась с цифры 1. Для этого еще раз войдите в раздел «Вставка — номер страницы». Выберите функцию «Формат номеров страниц...». Нумерация страниц должна начинаться с цифры 1. Нажмите ОК.

8) Основной текст доклада может содержать несколько озаглавленных частей и обязательно должен заканчиваться выводами (несколькими сформулированными вами предложениями, повторяющими наиболее важные

идеи работы) и списком использованной научной литературы и/или сетевых ресурсов (от 2 до 10 наименований). Назначьте заголовкам работы (в том числе фразе «Список научной литературы») стиль «Заголовок 1» (в разделе «Главная — Стили»).

9) В завершение работы над текстом выполните его проверку. Для этого нажмите вкладку «Рецензирование», выберите функцию «Правописание» и в открывшемся окне последовательно проверьте все слова и синтаксические конструкции, которые программа считает неправильными. Неправильные с точки зрения компьютерной программы слова выделены красным цветом, неправильные конструкции (слишком сложные предложения или предложения, в которых отсутствуют необходимые знаки препинания) — зеленым. Вы можете использовать в процессе проверки следующие функции:

   «Пропустить», если вы настаиваете на своем варианте написания,

   «Добавить», если слово, например, новый термин или фамилия ученого, распознается программой как неправильное, хотя оно встречается в работе несколько раз,

   «Заменить», если в слове допущена опечатка: в этом случае слово будет исправлено на предлагаемый программой вариант.

Если в предлагаемом программой списке вариантов для замены слова нет того варианта, который вам требуется, исправьте слово вручную.

Часть 2

ОБЛАСТИ ПРИМЕНЕНИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ЛИНГВИСТИКЕ

2.1. Автоматический анализ и синтез звучащей речи

Этапы автоматического анализа речи. Ввод в компьютер звучащей речи. Аналоговый и цифровой звуковой сигнал. Пословный и пофонемный анализ речи. Программы обработки звучащей речи и голосового управления компьютером. Методы автоматического синтеза речи.

Одним из первых важных шагов использования информационных технологий в лингвистике является дигитализация текстов — переведение языкового материала, существующего в печатном или устном виде, в цифровую форму. Именно в этом случае появляется возможность привлечения компьютеров для выполнения определенных операций над текстами на естественном языке: их преобразования, выделения их них отдельных элементов и создания (синтеза) аналогичных текстов.

В связи с принципиальными различиями в способах дигитализации и обработки звучащей речи и печатных текстов в нашей работе эти явления рассматриваются в разных параграфах. Первый параграф посвящен вопросам автоматической обработки и синтеза звучащей речи, а во всех последующих рассматриваются автоматические операции, производимые над печатными текстами.

При *автоматическом анализе* звучащей речи она преобразуется в печатный текст, над которым можно производить дальнейшие операции. *Автоматический синтез* звучащей речи представляет собой обратный процесс преобразования печатного текста, существующего в цифровой форме, в звучащий текст на естественном человеческом языке.

Процесс автоматического анализа речи включает следующие этапы:

1) ввод звучащей речи в компьютер с помощью микрофона,
2) выделение компьютерной программой в звуковом потоке отдельных знаков,
3) идентификация выделенных знаков звучащей речи со знаками языка.

Минимальными знаками звучащей речи являются звуки, производимые артикуляторным аппаратом человека. Каждый звук имеет свои акустические характеристики (высота, частота колебаний звуковых волн и т.д.), которые можно измерить специальными приборами (например, осциллографом).

Параметры звукового сигнала непрерывно меняются, и такой (непрерывный) тип сигнала называется аналоговым. В отличие от аналогового, цифровой сигнал представляет собой набор дискретных (отдельных) числовых значений, фиксирующих разные уровни звуковой волны. При использовании микрофона аналоговый звуковой сигнал преобразуется в аналоговый электрический, который с помощью аналогово-цифровых преобразователей, встроенных в звуковые карты современных компьютеров, переводится в дискретный цифровой сигнал [49].

Первые устройства автоматического распознавания устной речи, которых на сегодняшний день большинство, в качестве выделяемых в речевом потоке знаков использовали не звуки, а слова. Слова вводимой в компьютер речи идентифицировались со словами, заранее записанными диктором, читающим слова. Но такой тип распознавания речи связан с определенными ограничениями:

- *личность говорящего:* автомат распознает речь только определенного говорящего,
- *запас слов:* автомат распознает только ограниченное количество слов,
- *подготовленность речи:* автомат распознает речь, лишь если она подготовлена [25, 39].

Для преодоления этих ограничений требуется, чтобы компьютерная программа распознавала не слова, а звуки, т.е. работала не с дискретной речью (которая содержит паузы между словами), а со слитной естественной человеческой речью.

В основе пофонемного распознавания звуков речи лежит анализ 1) длительности и динамики звучания, 2) чередования акустического сигнала и пауз. При этом на основе универсальной классификации звуков Гуннара Фанта, Морриса Халле и Романа Якобсона акустические признаки звуков выводятся из артикуляционных. Правда, акустические признаки в отношении к артикуляционным оказываются недостаточно универсальными. Кроме того, в этой теории недостаточно учитывается слогоделение, акцентуация и ритм (главные носители смысла) [25, 39].

В настоящее время наиболее доступной формой точной фиксации звучащей речи (в том числе ее тембра и динамики) становится спектрограмма — фотографическое изображение звуков. Результаты наблюдений показывают, что в произнесении звуков активно используются четыре частоты называемые формантами. Так, на рис. 3. изображены форманты русских звуков *и* и *у*. При переходе от звука *и* к звуку *у* наиболее заметно изменение форманты F2 (рис. 3).

Задачей автоматического анализа звучащей речи при использовании спектрограмм становится перевод спектрограмм в фонологическую транскрипцию [25, 41].

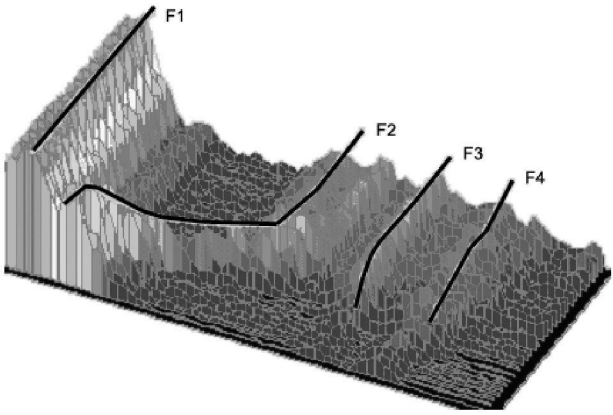


Рис. 3. **Спектрограмма русских звуков *и* и *у* [Фролов, Фролов]**

В итоге процесс автоматического анализа речи включает ввод слов в компьютер через микрофон, начитанных разными дикторами, их спектральную обработку и создание набора признаков, своеобразного образца слова, который выступает знаком языка. При распознавании звучащей речи реальные признаки составляющих ее единиц сравниваются с признаками и образцами слов, существующими в памяти машины. Результатом сравнения является транскрипция или орфографическая запись слова.

Но при автоматическом анализе слитной речи дополнительную трудность составляет отсутствие четких границ между словами. Человек для преодоления этой трудности кроме акустических сигналов обычно использует самые разные другие источники информации: ситуацию, контекст, структуру языкового высказывания, прошлый опыт в данной области и т.п. Аналогичные правила ученые пытаются применить и к машинам и стремятся задействовать в современных системах анализа речи кроме акустического другие уровни системы языка: лексический, синтаксический, семантический, прагматический.

Включение семантического уровня в автоматический анализ речи приводит, в частности, к следующим последствиям:

1) машина устанавливает, что введенные предложения многозначны и правдоподобны;

2) машина прогнозирует, что в определенных речевых контекстах могут возникать определенные типы общения; в зависимости от такого прогнозируемого типа общения машина интерпретирует предложение [35, 120].

Очевидно, что создание систем анализа речи такого сложного уровня предусматривает сотрудничество представителей самых разных специальностей. Для экономии времени и усилий ученых и практиков различные компании, в том числе *Microsoft*, выпускают средства анализа и синтеза речи в виде программных модулей и интерфейсов. Программисты, не обладающие познаниями в области лингвистики, математики и биологии, могут использовать готовые интерфейсы и программные модули в собственных разработках. Правда, в этом случае речевые возможности программ будут ограничены использованными средствами и технологиями. Например,

многие средства анализа и синтеза речи не способны работать с русским языком, что ограничивает их использование в России [49].

Можно назвать следующие примеры программ, в которых применяются средства автоматического анализа речи:

- программы голосового управления компьютером и бытовой техникой *VoiceNavigator* и *Truffaldino* (компания «Центр речевых технологий», С.-Петербург);
- комплекс голосового управления мобильным телефоном *DiVo* («Центр речевых технологий»);
- программный модуль *Voice Key* для идентификации личности по парольной фразе длительностью 3—5 секунд («Центр речевых технологий»);
- программы диктовки текста на английском языке: *VoiceType Dictation* (IBM), *DragonDictate* («Dragon Systems»); на русском языке: *Комбат* («Вайт Груп») и *Диктограф* («Voice Member Technology»);
- система распознавания речи, встроенная в *Microsoft Office XP* (работает только с английским языком);
- голосовой поиск (например, в поисковой системе *Google*).

Так, программа *VoiceNavigator* позволяет запускать компьютерные приложения и выполнять заданные команды голосом без использования клавиатуры. Перед применением программы ее необходимо обучить, произнеся в микрофон слова команд (команды можно произносить на любом языке и любым голосом). Чтобы программа начала распознавать голосовые команды, ее необходимо «разбудить», произнеся ключевое слово [49].

Использование модулей распознавания речи весьма перспективно в различных областях деятельности: в обслуживании клиентов, проведении судебных экспертиз, биометрии, обучении, научных исследованиях и т.д. Но массовое внедрение речевых технологий тормозится высокой стоимостью разработок и предлагаемых технологий, а также их пока еще низким качеством.

В целом задача автоматического анализа речи является весьма сложной и решена лишь отчасти. В сравнении с ней задача автоматического

синтеза речи оказывается более простой, и с примерами ее массового использования в обиходной жизни мы сталкиваемся постоянно. В частности, автоматически синтезируется речь в следующих ситуациях:

- называние текущего времени по телефону,
- объявление остановок в метро,
- называние остатка средств на счету и другие услуги мобильных операторов,
- оповещение систем гражданской безопасности и т.д.

Автоматический синтез (генерация) речи в настоящее время осуществляется путем составления слов и фраз из заранее записанных диктором образцов отдельных звуков (метод *компилятивного синтеза*) или путем моделирования речевого тракта человека (*формантно-голосовой* метод) [49].

Первый метод используется главным образом для синтеза относительно небольшого и заранее известного набора фраз. При этом обеспечивается довольно высокое качество звучания, поскольку синтезируемая речь базируется на элементах естественной человеческой речи. Тем не менее на стыке составляемых звуковых фрагментов возможны интонационные искажения и разрывы, заметные на слух. Кроме того, создание крупной базы данных звуковых фрагментов, учитывающей все особенности произношения фонем с разными интонациями, представляет собой сложную и кропотливую работу.

Второй метод оказывается более сложным, поскольку здесь необходимо точное моделирование особенностей речевого тракта человека, а также учет интонационной модуляции речи. В силу названных особенностей формантно-голосовая модель обладает относительно низкой точностью синтезируемых звуков речи.

В качестве примера программы, синтезирующей речь, можно назвать программу *Govorilka* (разработчик: А. Рязанов, бесплатная версия программы размещена по адресу http://www.vector-ski.com/vecs/govorilka). Основные особенности данной программы состоят в следующем:

- программа читает текст разными голосами и на разных языках, в том числе на русском;
- исходный текст для чтения может быть загружен из текстового файла или набран в окне программы при помощи клавиатуры;

- можно сохранить результаты синтеза речи, записав файл формата WAV или MP3.

Таким образом, несмотря на мощность современных компьютеров, проблема оснащения компьютера полноценным речевым интерфейсом еще далека от своего завершения. Главной проблемой при создании программ автоматического распознавания речи является то, что компьютер не умеет работать со смыслом. В синтезе речи уже имеются определенные достижения, которые внедрены в массовую практику.

Вопросы для обсуждения

1. Что такое знак? В чем различие между знаками языка и знаками речи?
2. В каких сферах ограничениями пословного распознавания звучащей речи можно пренебречь? Для каких сфер эти ограничения будут принципиально важными?
3. Какие артикуляционные признаки звуков вам известны?
4. Представители каких профессий должны быть задействованы в создании сложных систем анализа звучащей речи?

Рекомендуемая литература

1. Алексеев В. Услышь меня, машина // Компьютерра. 1997. № 49. http://offline.computerra.ru/1997/226/938
2. Марчук Ю.Н. Компьютерная лингвистика: учеб. пособие. М.: АСТ Восток — Запад, 2007. С. 38—44.
3. Мыркин В.Я. Введение в языкознание: учебник. Архангельск: Поморский университет, 2005. С. 57—69.
4. Фролов А.В., Фролов Г.В. Синтез и распознавание речи. Современные решения. 2003. http://frolov-lib.ru/books/hi/index.html

Лабораторная работа 4

Запись и обработка звуковых файлов

1. Подсоедините микрофон к компьютеру. Нажмите «Пуск» — «Все программы» — «Стандартные» — «Развлечения» — «Звукозапись». Откроется окно программы «Звук — Звукозапись», встроенной в операционную

систему MS Windows. Программа сохраняет и обрабатывает файлы только в формате звукозаписи .WAV.

2. В меню «Файл» выберите команду «Создать». Нажмите кнопку «Запись» (обозначена красным кружком) и произнесите в микрофон заранее подготовленный текст: одно предложение на русском и одно — на иностранном языке. Между предложениями сделайте небольшую паузу. Для завершения записи нажмите кнопку «Стоп» (обозначена прямоугольником). Сохраните запись со стандартными настройками, нажав для этого меню «Файл» и выбрав вкладку «Сохранить как...». В качестве имени файла введите вашу фамилию.
3. Проверьте запись: для этого нажмите кнопку «Воспроизвести», обозначенную треугольником. Запомните, на какой секунде записи заканчивается первое предложение.
4. Удалите часть записи — предложение на иностранном языке. Для этого переместите бегунок в точку файла, начиная с которой требуется удалить звукозапись. Для более точного поиска ориентируйтесь на счетчик времени, показывающий секунды и доли секунд. В меню «Правка» выберите команду «Удалить после текущей позиции». Сохраните оставшуюся часть записи («Файл» — «Сохранить как...»), выбрав в качестве имени файла начало русского предложения.
5. Откройте исходный файл, обозначенный вашей фамилией. Удалите из записи предложение на русском языке, переместив бегунок в точку, находящуюся между записанными предложениями, и выбрав из меню «Правка» команду «Удалить до текущей позиции». Сохраните запись («Файл» — «Сохранить как...»), выбрав в качестве имени файла начало предложения на иностранном языке.
6. Вставьте один записанный файл в другой. Для этого откройте файл, обозначенный вашей фамилией. Поместите бегунок в конец файла. В меню «Правка» выберите команду «Вставить файл». Дважды щелкните по файлу, обозначенному началом русского предложения. Проверьте запись и убедитесь, что в ней скомбинированы два исходных звуковых файла.
7. Поработайте с меню «Эффекты»: выберите пункт «Увеличить громкость (на 25%)» и «Добавить эхо». Проверьте запись. Сохраните результат обработки, назвав его следующим образом: Ваша фамилия_Эффекты.

Лабораторная работа 5

Автоматический синтез устной речи

1. Перейдите по ссылке http://www.bloxpot.net/2010/ 05/blog-post_29.html. Просмотрите видеосюжет о возможностях автоматического синтеза рече-

вых технологий. Прослушайте записи фраз, синтезированных в различных программах. Оцените качество синтезированной речи каждой программы.

2. Перейдите по ссылке http://text-to-speech.imtrans lator.net. Введите в диалоговое окно одно или несколько предложений на русском, английском и других известных вам языках. Прослушайте вариант озвучивания этих фраз, предлагаемых программой. Для каких целей можно использовать данную программу?
3. Перейдите по ссылке http://rssradio.ru. Протестируйте различные возможности автоматического озвучивания новостей российских интернет-порталов. Насколько полезной вы считаете функцию автоматического озвучивания новостей?
4. Перейдите по ссылке http://mp3book2005.ru/3.htm. Прослушайте примеры аудиозаписей книг, предлагаемые на сайте. Оцените возможности использования программы.
5. Составьте перечень недостатков автоматического синтеза речи, выявленных вами на материале рассмотренных программ. В какой из программ этих недостатков меньше всего?

2.2. Автоматическое распознавание текста

Ввод печатного текста в компьютер. Распознавание текста с помощью OCR-программ.

Автоматический анализ печатного текста, также как и анализ звучащей речи, начинается с его ввода в компьютер. Поскольку современный человек имеет дело главным образом с информацией, размещенной на печатных носителях, остановимся на процессах автоматической обработки печатных текстов подробнее и посвятим им несколько отдельных параграфов. В первом из этих параграфов рассмотрим ввод печатного текста в компьютер и связанный с этим процесс распознавания печатного текста.

Для ввода информации в компьютер используются специальные устройства — клавиатура, мышь и др., но наиболее удобным инструментом для ввода большого количества печатных текстов является сканер.

Сканер — это устройство ввода, работающее по принципу фотоаппарата, т.е. позволяющее компьютеру «увидеть» текст в виде фо-

тографии [20, 53]. Чтобы компьютер смог «понять» этот текст, т.е. перевести графическое (растровое) изображение символов в текстовую форму, при которой у каждого символа имеется свой двоичный код (например, в системе кодировок ASCII), требуется программа автоматического распознавания символов (англ. OCR = *Optical Character Recognition*).

Символами являются любые буквы, знаки препинания и другие знаки текста (апостроф, кавычки, тире, скобки и т.д.). Слово понимается как последовательность символов между двумя соседними пробелами.

Программа автоматического распознавания текста (OCR-программа) — это компьютерная программа, позволяющая преобразовать текст с бумажного носителя в электронный текстовый файл, который в дальнейшем может обрабатываться человеком в любом текстовом редакторе [20, 53]. Такие программы обычно предлагаются с каждым приобретаемым сканером, но наиболее известными и полифункциональными являются OCR-программы *FineReader* (компании *Abbyy*) и *CuneiForm* (фирмы *Cognitive Technologies*).

С другими программами автоматического распознавания текстов можно познакомиться, например, в интернет-ресурсе, размещенном по адресу http://kompkimi.ru/?p=617 (дата обращения: 02.02.2012).

Результат распознавания большинством OCR-программ весьма точен, хотя некоторые трудности в распознавании текста приводят к ошибкам, которые впоследствии приходится исправлять вручную. Трудности распознавания могут быть вызваны следующими особенностями печатного текста [20, 53—54; 6, 34]:

- использование шрифта разной гарнитуры и размера,
- использование в тексте нескольких языков,
- размещение текста в несколько колонок,
- включение в текст таблиц и рисунков,
- искажения символов (разрывы, слипания букв и т.п.),
- посторонние включения в изображение и т.д.

Названные трудности решаются, если дополнить системы автоматического распознавания текстов возможностью работы машины

со смыслом документа, т.е. вывести OCR-программы на более высокий уровень искусственного интеллекта.

В целом точность распознавания OCR-программ на текстах хорошего и среднего качества достигает 99%, что позволяет считать проблему массового ввода печатных текстов в компьютер практически решенной.

Вопросы для обсуждения

1. Охарактеризуйте основные возможности OCR-программ.
2. Каковы перспективы развития OCR-программ?
3. Что такое «интеллектуальное распознавание»?
4. Охарактеризуйте особенности одной из систем автоматического распознавания текста.

Рекомендуемая литература

1. Башмаков И.А, Башмаков А.И. Интеллектуальные информационные системы. М.: МГТУ им. Н.Э. Баумана, 2005. С. 32—40.
2. Всеволодова А.В. Компьютерная обработка лингвистических данных: учеб. пособие. 2-е изд., испр. М.: Флинта; Наука, 2007. С. 47.
3. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 53—55.
4. Сканирование и распознавание: профессиональный подход: материалы курса дистанционного обучения. www.online-academy.ru/scan.htm (дата обращения: 02.02.3012).

Лабораторная работа 6

Сканирование текстового документа

Для сканирования возьмите текст любого печатного издания (книги, журнала) на ваш выбор. Нужно отсканировать три страницы (разворота) печатного издания.

1. Откройте программу ABBYY FineReader (через меню «Пуск» — Все программы).
2. Откройте вкладку меню «Файл» — «Сканировать несколько страниц...» или нажмите кнопку «Сканировать» над основным рабочим полем программы.

Положите в сканер печатный текст. Следуйте инструкциям в открывшемся окне работы со сканером. После предварительного сканирования при необходимости поверните изображение. Сканируйте первую страницу текста, вложенного в сканер, нажав кнопку «Scan» окна работы со сканером. Отсканированная страница должна появиться в программе ABBYY FineReader под номером 1.

3. Повторите процедуру сканирования со страницами 2 и 3. После того как эти страницы появились в программе ABBYY FineReader под номерами 2 и 3, закройте окно работы со сканером и продолжайте работу только с программой ABBYY FineReader.
4. Проверьте язык распознаваемого документа. Он должен соответствовать языку того документа, который вы сканируете. При необходимости выберите нужный язык распознавания из предлагаемого списка.
5. Нажмите кнопку «Распознать все» над основным рабочим полем программы ABBYY FineReader.
6. После окончания процесса распознавания выберите функцию «Передать все в MS Word» (нажав вкладку меню файл — передать все... или нажав треугольник рядом с кнопкой MS Word над основным рабочим полем программы ABBYY FineReader.
7. Сохраните получившийся файл MS Word под названием Л4_Номер группы_Ваша фамилия, например, Л4_10_Иванов. Добавьте в начале документа библиографические данные книги: Фамилия и инициалы автора. Название книги. Город: Издательство, год. Кол-во страниц по образцу: *Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. М.: Академия, 2004. 208 с.*

2.3. Автоматическое аннотирование и реферирование текста

Понятие автоматического аннотирования и реферирования текста. Виды рефератов. Примеры систем автоматического аннотирования.

В условиях все возрастающего количества текстов в окружающем человека мире возникает проблема: как в этом море информации найти нужные документы и познакомиться с их содержанием? Решению данной проблемы может помочь составление рефератов и аннотаций полнотекстовых документов. Они дают читателю пред-

ставление о содержании исходных документов и позволяют оценить степень необходимости обращения к полным текстам каждой работы. Кроме того, рефераты и аннотации акцентируют внимание читателя на новых сведениях, т.е. позволяют за небольшой промежуток времени узнать много новой информации.

Рефераты и аннотации составляются вручную, например самим автором исходного текста или библиографическим работником, или автоматически, с помощью специальных компьютерных программ. Наиболее качественным является первый вид рефератов и аннотаций, поскольку в этом случае создается новый текст, называющий основную мысль высказывания и отличающийся связным характером. Но для обработки большого массива текстов за минимальное количество времени требуется привлечение автоматических средств для решения задачи реферирования и аннотирования текстов.

Реферат определяется как связный текст, который кратко выражает

- центральную тему,
- предмет
- цель,
- методы,
- результаты исследования [20, 55].

Рефераты обычно составляют к научно-техническим документам: научным монографиям, статьям, патентам на изобретение и др. В зависимости от жанра исходного текста (монография, статья, патент и др.) и от предметной области (медицина, химия, лингвистика и т.д.) заданные элементы реферата могут различаться. Так, для научных рефератов дополнительно к названным выше элементам реферата прибавляется краткое изложение сути, практической апробации и перспектив исследования [8, 93—94].

Различают следующие виды рефератов [8, 89]:

- *связный текст* — новое текстовое образование, порождаемое на основе логико-смыслового анализа исходного текста;
- *реферат-клише* — модификация заданной клишированной структуры, пустые ячейки которой заполняются после анализа заданного текста;

- *квазиреферат* — перечень наиболее информативных предложений текста.

Очевидно, что для автоматического создания рефератов — связных текстов требуются более сложные компьютерные программы, чем для создания рефератов-клише и квазирефератов.

Некоторые исследователи считают реферат и аннотацию синонимами [8, 88], а некоторые предлагают разводить эти понятия, определяя аннотацию как краткое изложение содержания документа, дающее общее представление о его теме [20, 55]. Согласно этому определению в отличие от реферата, знакомящего читателя с сутью излагаемого в документе содержания, аннотация выполняет лишь сигнальную функцию (есть публикация на определенную тему).

В большинстве программ, направленных на автоматическое составление краткого содержания текста, можно задать разную степень компрессии текста, т.е. одна и та же программа создает как развернутые рефераты, так и краткие аннотации. В связи с этим в отношении автоматического процесса составления краткого содержания текста обычно используется двойное обозначение: автоматическое реферирование и аннотирование текста.

Создаваемые в процессе реферирования и аннотирования аннотации и рефераты представляют собой вторичные документы. Первичными (или исходными) документами являются сами книги, статьи, патенты и др.

Программы автоматического аннотирования и реферирования ориентированы на то, как это делает человек. Для человека этот процесс включает следующие этапы [20, 56—57]:

1) подготовительный: определение темы текста, его понимание;
2) аналитический: деление текста на фрагменты (абзацы и т.п.) и выделение в каждом фрагменте главных смысловых слов, которые составляют план будущего реферата;
3) непосредственное составление реферата или аннотации: соединение выделенных смысловых единиц в связный текст.

Главными смысловыми единицами исходного текста выступают ключевые слова, ключевые словосочетания и ключевые предложения. *Ключевое слово* — знаменательное слово, относящееся к основ-

ному содержанию текста и повторяющееся в нем несколько раз. *Ключевое словосочетание* — сочетание слов, среди которых есть одно или несколько ключевых. Ключевое предложение — предложение, которое содержит несколько (два и более) ключевых слов [20, 57].

По способам выделения из исходных текстов ключевых словосочетаний и предложений различаются следующие методы автоматического реферирования и аннотирования текстов [20, 59]:

1) статистические,
2) позиционные,
3) логико-семантические.

При *статистическом методе* принадлежность слова к категории ключевых определяется его статистическими характеристиками: ключевое слово согласно этому методу встречается среди знаменательных слов текста наибольшее количество раз. Ключевое предложение, соответственно, содержит несколько ключевых слов, которые располагаются на небольшом расстоянии друг от друга.

В *позиционном методе* принцип отнесения предложения к ключевым опирается на его местонахождение в тексте: ключевые предложения входят в заголовок, подзаголовок, находятся в начале и конце текста.

Целью *логико-семантического метода*, при котором учитывается структура и семантика текста, является выделение предложений с наибольшим функциональным весом. Такими предложениями считаются те, которые содержат семантически значимые слова, особым образом связаны с другими предложениями, имеют определенный синтаксический тип предложения и т.п.

Наиболее простыми системами автоматического реферирования и аннотирования является функция *AutoSummarize* в *MS Word*, системы *Intelligent Text Miner*, *Oracle Context* и *Inxight Summarizer* (компонент поискового механизма *AltaVista*) (IBM). Правда, возможности этих программ ограничены выбором оригинальных фрагментов из исходного документа и их соединением в короткий текст [50].

Кроме того, можно привести примеры следующих систем автоматического реферирования и аннотирования текстов:

- ОРФО 5.0 (компания «Информатик»): программа включает функцию автоматического аннотирования русских текстов;
- «Либретто» (компания «МедиаЛингва»): программа встраивается в Word и обеспечивает автоматическое реферирование и аннотирование русских и английских текстов;
- поисковая система «Следопыт», которая включает средства автоматического реферирования и аннотирования документов;
- программы Extractor и TextAnalyst (компания «Медиасистемы»), которые выдают последовательности именных групп, выделенных с помощью синтаксических анализаторов.

В целом можно констатировать, что автоматические рефераты и аннотации представляют собой, по сути, квазирефераты, т.е. результатом автоматической компрессии текста в большинстве случаев становится либо набор ключевых слов, либо перечень ключевых предложений, что, впрочем, в значительной степени помогает решить задачу аннотирования и реферирования большого объема текстов в малые сроки.

Вопросы для обсуждения

1. Опишите этапы составления реферата текста.
2. Представьте известные вам системы автоматического реферирования и аннотирования текстов.
3. Какие задачи являются перспективными для систем автоматического реферирования и аннотирования текстов?

Рекомендуемая литература

1. Башмаков И.А, Башмаков А.И. Интеллектуальные информационные системы. М.: МГТУ им. Н.Э. Баумана, 2005. С. 77—90.
2. Беляева Л.Н. Лингвистические автоматы в современных гуманитарных технологиях: учеб. пособие. СПб.: Книжный Дом, 2007. С. 87—101.
3. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 55—75.

Лабораторная работа 7

Использование функции «Автореферат» *MS Word*

1. Откройте любой текстовый документ в *MS Word 2007*. В верхней строке командного меню нажмите вкладку «Настройка панели быстрого доступа» (треугольник, обращенный вниз). Выберите вкладку «Другие команды».
2. В открывшемся окне найдите указатель «Выбрать команды из...», в котором нужно установить параметр «Команды не на ленте». Выделите в списке команд «Автосуммирование» и нажмите «Добавить». Далее нажмите «ОК» и убедитесь, что значок «Автосуммирование» появился в верхней строке командного меню.
3. Нажмите значок «Автосуммирование», выберите вкладку «Автореферат». Выберите вид и размер реферата, нажмите «ОК».
4. Проанализируйте получившийся реферат и отразите результаты анализа в таблице.

Параметр анализа	**Ваш комментарий**
Связный текст или набор словосочетаний/предложений	
Функциональная нагруженность элементов реферата	
Отражены ли необходимые структурные компоненты реферата (тема, цель, методы, результаты работы)	
Общий вывод	

5. Составьте вручную реферат того же самого исходного текста. Ориентируйтесь при этом на этапы составления реферата, названные в теоретической части раздела.
6. Сохраните оба реферата (автоматический и созданный вами) и таблицу с комментарием об автоматическом реферате (задание 4) в *Word* под названием Л7_Номер группы_Ваша фамилия, например, Л7_10_Иванов.

2.4. Автоматический анализ и синтез текста

Графематический, морфологический, синтаксический и семантический анализ текста. Понятие токенизации, парсера. Формальная грамматика. Машинная основа, машинное окончание. Автоматический синтез текста.

Автоматический анализ текста включает ряд весьма сложных операций, которые компьютер выполняет над текстом на естественном

человеческом языке согласно заданному алгоритму. При *автоматическом анализе* текст последовательно преобразуется в его лексемно-морфологические, синтаксические и семантические представления, понятные компьютеру [13, 14]. Обратный процесс преобразования лексемно-морфологических, синтаксических и семантических компьютерных представлений в текст на естественном языке называется *автоматическим синтезом текста*.

Автоматический анализ и синтез текста являются важными задачами компьютерной лингвистики как с точки зрения развития теории (разработки лингвистических основ создания искусственного интеллекта), так и с точки зрения реализации практических нужд человека, например, создания эффективных систем машинного перевода.

Автоматический анализ текста включает ряд этапов [1, 94, 106—107]:

1) *графематический анализ*: выделение границ слов, предложений, абзацев и других элементов текста (например, врезок в газетном тексте);
2) *морфологический анализ*: определение исходной формы каждого использованного в тексте слова и набора морфологических характеристик этого слова;
3) *синтаксический анализ*: выявление грамматической структуры предложений текста;
4) *семантический анализ*: определение смысла фраз.

Графематический анализ определяется также как токенизация (от англ. *token* = отдельное слово, фраза или любой другой значимый элемент текста[1]). Формальными сигналами границ текстовых элементов выступают разделители различного рода: пробелы, обозначающие границы между словами, *прописные буквы и знаки препинания*, обозначающие границы между предложениями и составными частями предложений, *абзацные отступы*, обозначающие границы между связанными по смыслу группами предложений и т.п. [7, 79].

Однако формальный метод определения границ слов применим не всегда. Например, в китайском языке нет формальных границ

[1] http://en.wikipedia.org/wiki/Tokenization

слов [54]. Кроме того, даже в распространенных европейских языках существуют устойчивые сочетания слов, разделенные пробелом, которые следует воспринимать как одну лексему, например, New York. Очевидно, что такие случаи следует учитывать в системах графематического анализа, например, путем создания списков многословных лексем.

При **морфологическом анализе** каждое использованное в тексте слово возводится к его исходной форме и определяется набор морфологических характеристик текстовой формы слова: часть речи; род, число и падеж для существительных, число и лицо для глаголов и т.п.

Каждое употребленное в тексте слово называется словоформой (или словоупотреблением). Для обеспечения связности текста требуется повтор тех же самых слов, поэтому нередко разные словоформы одного или нескольких предложений текста возводятся к одной и той же исходной форме, ср.:

Вот моя деревня;
Вот мой дом родной.
Вот качусь я в санках
По горе крутой (И.З. Суриков).

Алфавитно-частотный словарь словоформ этого фрагмента стихотворения выглядит так: *в* — 1, *вот* — 3, *горé* — 1, *деревня* — 1, *дом* — 1, *качусь* — 1, *крутой* — 1, *мой* — 1, *моя* — 1, *по* — 1, *родной* — 1, *санках* — 1, *я* — 1. Кроме неизменяемой частицы вот, употребленной 3 раза, отмечаем также притяжательное местоимение 1-го лица ед. числа, употребленное в формах *мой* и *моя*.

В привычных нам словарях обычно перечисляются не словоформы, а слова, приведенные к определенной исходной форме. В качестве такой исходной формы употребленных в тексте словоформ в зависимости от типа языка может служить *лемма* (словарная форма лексемы) или основа (ядерная часть слова без словоизменительных морфем). Например, английские словоформы *swim*, *swims*, *swam* и swimming восходят к одной лемме *swim*.

Во флективных и агглютинативных языках с богатым словоизменением для сохранения всех возможных словоформ потребуются достаточно значительные ресурсы памяти. Например, русское суще-

ствительное, изменяющееся по числам (2 числа) и падежам (6 падежей), имеет 12 словоформ. Русский глагол характеризуется еще более сложным набором грамматических характеристик и соответственно имеет достаточно значительное количество словоформ [20, 83]. В этом случае в качестве исходной формы, к которой возводится слово, удобнее использовать его основу.

Правда, в морфологическом анализе термин «основа» не всегда имеет тот же смысл, который вкладывается в него в канонической (школьной) грамматике. Например, если в слове встречается чередование букв (*сидеть — сижу*, *друг — друзья* и т.п.), то основой (точнее, квазиосновой, или машинной основой) в этих случаях выступает часть слова не только без словоизменительных морфем, но и без чередующихся букв, т.е. *си#* и *дру#*, соответственно.

Такой тип выделения основ получил название *стемминга*, т.е. возведения разных словоформ к одной квазиоснове. Стемминг вполне подходит для решения некоторых автоматических задач, например, для осуществления поиска в Интернете. Так, пользовательскому запросу *фотографи* в качестве полной или неполной квазиосновы соответствуют существительное фотография и прилагательное фотографический. В результате поиска пользователь получит список документов со словосочетанием *фотографический портрет* и со словосочетанием *портретная фотография* [46].

Для морфологического анализа важно не только понятие *машинной основы*, понимаемой как последовательность букв от начала словоформы, общая для всех словоформ, входящих в формообразовательную парадигму данного слова. Следующий шаг — это определение частеречной принадлежности слова (частеречный тегинг) и его морфологических характеристик, что чаще всего происходит с опорой на словоизменительные элементы слова (машинные окончания).

Машинные окончания — элементы, описывающие формоизменение конкретной лексемы и представляемые в виде парадигм. Все возможные наборы машинных окончаний зафиксированы в *типовой парадигме* лексемы. При этом, с одной стороны, можно наблюдать совпадения типовых парадигм (и, соответственно, машинных окончаний) разных лексем, например, *ручка* и *кочка*, а с другой, совпадения машинных основ лексем, имеющих разные типовые парадигмы,

ср. типовые парадигмы машинной основы *лож#*, относящейся к лексемам *ложь* и *ложиться* [8, 144—145].

По машинным окончаниям, входящим в определенные типовые парадигмы, осуществляется полная морфологическая характеристика каждой словоформы, например:

Девочка {девочка = S, жен, од = им, ед}

мыла {мыть = V, несов = прош, ед, изъяв, жен, перех | мыло = S, сред, неод = им, мн | = S, сред, неод = род, ед | = S, сред, неод = вин, мн}

пол {пол = S, муж, неод = им, ед | = S, муж, неод = вин, ед | = A, кратк, муж, им, ед}.

В приведенном анализе можно увидеть лексико-морфологическую многозначность второго и третьего слова. Выбор правильной формы осуществляется человеком с учетом синтаксической роли слова в предложении и его смысла. Автоматическое разрешение многозначности или *снятие омонимии*, понимаемое как выбор правильной интерпретации словоформы, допускающей несколько вариантов толкований, происходит путем ручной разметки или автоматически, на основе вероятностных моделей (например, в английском языке наиболее вероятно сочетание неопределенного артикля и существительного, следующего за ним) или на основе правил, созданных автоматически или человеком. Примеры таких правил следующие:

- Если словоформа может быть как глаголом, так и существительным, и перед ней стоит артикль, эта словоформа в данном случае является существительным.
- Если словоформа может быть как предлогом, так и подчинительным союзом, и если после нее до конца предложения нет глагола, эта словоформа в данном случае является предлогом [46].

Для автоматического морфологического анализа применяются *парсеры* — специальные компьютерные программы для автоматического анализа слов [32]. Кроме морфологических существуют и синтаксические парсеры, применяемые для автоматического анализа синтаксических структур предложений.

В целом морфологический анализ включает в себя следующие этапы [46]:

1) нормализация словоформ, имеющая вид лемматизации, т.е. сведения различных словоформ к некоторому единому представлению — к исходной форме (лемме) или стемминга, т.е. возведения разных словоформ к одной квазиоснове;
2) частеречный тэгинг, т.е. указание части речи для каждой словоформы в тексте;
3) полный морфологический анализ — приписывание грамматических характеристик словоформе.

При **синтаксическом анализе** необходимо определить роли слов в предложении и их связи между собой. Результатом этого этапа автоматического анализа является представление синтаксических связей каждого предложения в виде моделей, например в виде дерева зависимостей.

Проблемой синтаксического анализа выступает наличие альтернативных вариантов синтаксического разбора (синтаксической многозначности), ср.:

три пальто → (сколько?) *три* (чего?) *пальто*

три пальто → (что делай?) *три* (что?) *пальто*

Возникновение синтаксической многозначности обусловливается лексико-морфологической многозначностью словоформ (одна и та же словоформа может восходить к различным исходным формам или к разным морфологическим формам одной лексемы), а также неоднозначностью самих правил разбора, которые могут иметь целью представление синтаксической структуры, например, в виде дерева непосредственных составляющих или дерева зависимостей. Так, предложение «Девочка мыла пол» описывается в первом случае моделью, представленной на рис. 4, а во втором — рис. 5.

В модели непосредственно составляющих важно разбиение синтаксической структуры на пары ее элементов: предложение (S) разбивается на группу подлежащего (NP), представленную в данном случае одним существительным (N), и группу сказуемого (VP). Вторая делится на изменяемый глагол (V) и дополнение (O). В дереве

зависимостей исходным пунктом анализа выступает сказуемое (V), находящееся в вершине графа, от которого зависят подлежащее (N) и дополнение (O). В итоге в обоих типах анализа выделяются одни и те же синтаксические единицы — N, V и O — но синтаксические отношения между ними оказываются разными.

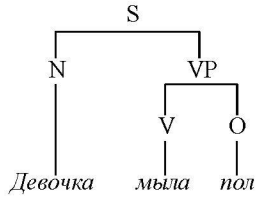


Рис. 4. **Дерево непосредственно составляющих**

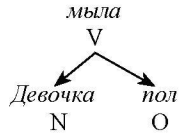
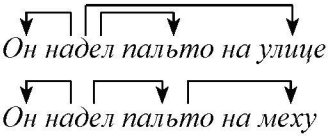


Рис. 5. **Дерево зависимостей**

Правда, в некоторых случаях на первый взгляд идентичная синтаксическая структура требует построения разных синтаксических моделей, ср. [3, 251]:



Чтобы выбрать правильную модель, отражающую синтаксические отношения в конкретном предложении, в подобных случаях необходимо привлечь семантику.

Семантический анализ представляет собой, пожалуй, наиболее сложное направление автоматического анализа текста. В этом случае требуется установление семантических отношений между словами в тексте, объединение различных языковых выражений, относящихся к одному и тому же понятию, и т.п.

Для семантического анализа предложений используются падежные грамматики и семантические падежи (валентности). В этом случае семантика предложения описывается через связи главного слова (глагола) с его семантическими актантами. Например, глагол *передать* описывается семантическими падежами дающего (агенса), адресата и объекта передачи [11, 96].

В основе семантического анализа лежит утверждение о том, что значение слова не является элементарной семантической единицей.

Оно делится на более элементарные смыслы — единицы словаря семантического языка. Эти единицы семантического языка являются своеобразными атомами, из различных комбинаций которых складываются «молекулы» — значения реальных слов естественного языка [3, 254].

Например, если имеются элементарные смыслы «сам», «кто-то», «иметь», «заставлять», «переставать», «начинать» и «не», то с их помощью мы можем определить целую группу слов русского языка. Кроме семи названных слов, являющихся одновременно и элементами семантического языка, и словами русского языка, сюда относятся слова: 1) владеть = «иметь», 2) обладать = «иметь», 3) брать = «заставлять себя иметь», 4) давать = заставлять кого-то иметь» и т.д. [там же].

Именно семантический анализ позволяет решить проблемы многозначности (омонимии), возникающей при автоматическом анализе на всех языковых уровнях.

- Лексическая омонимия: совпадение звучания и/или написания слов, не имеющих общих элементов смысла, например, рожа — лицо и вид болезни.
- Морфологическая омонимия: совпадение форм одного и того же слова (лексемы), например, словоформа пол соответствует именительному и винительному падежам существительного пол.
- Лексико-морфологическая омонимия (наиболее частый вид омонимии): совпадение словоформ двух разных лексем, например, мыла — глагол мыть в единственном числе женского рода прошедшего времени и существительное мыло в единственном числе, родительном падеже.
- Синтаксическая омонимия: неоднозначность синтаксической структуры, имеющей несколько интерпретаций, например: *Эти типы стали есть в цехе* (словоформа *стали* может интерпретироваться как существительное или как глагол), *Flying planes can be dangerous* (известный пример Хомского, в котором словоформа *Flying* может интерпретироваться либо как прилагательное, либо как существительное) [11, 93—94].

Автоматический синтез представляет собой процесс производства связного текста, отдельные этапы которого являются теми же, что и при морфологическом анализе, но применяются в обратном порядке: сначала осуществляется семантический синтез, затем синтаксический, морфологический и графематический.

Семантический синтез представляет собой переход от смысловой записи фразы к ее синтаксической структуре; синтаксический — переход от синтаксической структуры фразы к представляющей фразу цепочке лексико-грамматических характеристик словоформ; *лексико-морфологический* — переход от лексико-грамматической характеристики к реальной словоформе [27]. При морфологическом синтезе по нормальной форме слова и его параметрам программа находит соответствующую словоформу. Графематический синтез объединяет слова в единый текст, следит за соответствием фрагментов входного текста фрагментам выходного [21, 108].

Одной из первых компьютерных программ, синтезирующих письменный диалог на английском языке, явилась программа американского ученого Джозефа Вейценбаума «Элиза». Ее первый вариант был апробирован в 1966 году. Эта программа имитировала диалог с психотерапевтом, реализуя технику активного слушания. Программа была названа в честь Элизы Дулитл, героини из пьесы «Пигмалион» Бернарда Шоу, которую обучали языку «высшего класса людей». По сути, целью программы было не моделирование мышления в точном смысле, а моделирование речевого поведения, что было обусловлено ограниченными программными ресурсами, а также начальным уровнем лингвистического анализа и синтеза [4, 21].

Программа включала минимум лингвистической информации: 1) комплекс ключевых слов, которые актуализируют некоторые устойчивые речевые формулы, 2) способность трансформировать предыдущее высказывание в общий вопрос [там же]. Иллюстрацией работы программы «Элиза» выступает следующий фрагмент общения молодой женщины и программы «Элиза», опубликованный ее создателем Д. Вейценбаумом в 1966 г. Реплики, смоделированные программой, выделены прописными буквами.

Men are all alike.

IN WHAT WAY?

They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE?
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY?
I need some help, that much seems certain [цит. по: 57].

Несмотря на простоту алгоритмов, использованных в создании программы, с ее помощью практически удалось опровергнуть известный тест английского ученого Алана Тьюринга, предложенный в 1950 г. Сущность *теста Тьюринга* формулируется следующим образом: проблема создания искусственного интеллекта решена, если большинство участников общения не смогут установить, с кем они беседуют — с человеком или машиной. При проведении экспериментов с программой «Элиза» 62% испытуемых, которым было предложено определить, с кем они общаются, решили, что им отвечал человек [4, 21].

На современном этапе системы компьютерного моделирования письменных диалогов на естественном языке используют более сложные алгоритмы. В частности, создан специальный язык разметки для искусственного интеллекта AIML (*Artificial Intelligence Markup Language*), используемый для создания виртуальных агентов (или ботов). Боты, моделирующие диалог с собеседником, используются в компьютерных играх и на корпоративных веб-страницах, например, для ответов на вопросы пользователя о возможностях мобильного оператора или торговой сети.

Вопросы для обсуждения

1. Назовите и кратко охарактеризуйте уровни естественного языка, релевантные для морфологического анализа и синтеза текста.
2. Дайте определения основным понятиям автоматического анализа текста: *слово*, *словоформа*, *лемма*, *машинная основа*, *стемминг*, *частеречный тэгинг*, *парсер*, *тест Тьюринга*.

3. Назовите и дайте краткую характеристику этапам автоматического анализа текста.
4. Назовите и дайте краткую характеристику этапам автоматического синтеза текста.
5. Охарактеризуйте системы компьютерного моделирования диалогов, в том числе роботы-автоответчики. Как происходит обучение роботов? Как распознать робот-автоответчик?

Рекомендуемая литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. С. 91—97, 106—111.
2. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. 3-е изд. М.: ЛКИ, 2007. С. 20—25.
3. Всеволодова А.В. Компьютерная обработка лингвистических данных: учеб. пособие. 2-е изд., испр. М.: Флинта: Наука, 2007. С. 50—51, 66—67.
4. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 33—34.

Упражнения

1. Определите машинные основы следующих слов:

Слово	**Машинная основа**	**Слово**	**Машинная основа**
семья		*окружать*	
весна		*сидеть*	
дело		*друг*	
пианино		*ты*	
стремительный		*желтый*	
развивать		*петь*	

2. Приведите примеры совпадения машинных основ для слов русского языка и запишите их ниже:
   - основе *лож#* соответствуют типовые парадигмы для описания слов *ложь* и *ложиться*;
   - _____;
   - _____.

3) Определите исходную форму выделенных словоформ в примере ниже. Какие виды омонимии иллюстрирует данный текст?

Я траву ***косил косой***,
Дождик вдруг пошел ***косой***.
Бросил я тогда ***косить***
И на Стешу стал ***косить***.
Ну а Стеша, ох, краса,
Как огонь её ***коса****!* (www.stihi.ru/2011/06/03/8515)

4) Выберите текст одного из СМИ на иностранном языке объемом до 1 страницы и составьте список слов для словаря словоформ. Укажите частотность появления каждой словоформы в проанализированном тексте.

5) Опишите через элементарные семантические смыслы «сам», «кто-то», «иметь», «заставлять», «переставать», «начинать» и «не» слова русского языка: *приобретать, утрачивать, сохранять, лишать*.

6) Определите, какие фразы во фрагменте диалога ниже — собеседника А или Б — смоделированы машиной. Приведите аргументы в пользу вашего выбора.

А *Говорите!*
Б *Меня волнует погода.*
А *Как давно вас волнует погода?*
Б *С тех пор, как умерла моя мать.*
А *Расскажите больше о Вашей семье.*
Б *Они все в области компьютеров.*
А *Машины пугают вас.*
Б *Нет* [пример цит. по: 4, 22].

Лабораторная работа 8

Автоматический анализ текста

1. Посетите сайт http://starling.rinet.ru/morph.htm и протестируйте работу онлайн-анализатора морфологии. Опишите результаты анализа в таблице для различных частей речи русского и английского языков.

Слово	**Пример, рус.**	**Пример, англ.**	**Ваш комментарий о возможностях морфологического анализатора**
Существительное			
Глагол			
Прилагательное			
Наречие			

2. Посетите сайт http://sz.ru/parser. Введите 3 любые предложения на русском языке, имеющие разную синтаксическую структуру. Сравните результаты их синтаксического анализа в таблице, оценивая при этом полезность представленной в анализе лингвистической информации. Прокомментируйте возможности применения подобных систем анализа.

Предложение, рус.	**Лингвистическая информация, представлена в синтаксическом анализе**
1.	
2.	
3.	
Общий комментарий	

3. Посетите сайт http://nlp.stanford.edu:8080/parser/index.jsp. Введите 3 любые предложения на английском языке, имеющие разную синтаксическую структуру. Сравните результаты их синтаксического анализа в таблице, оценивая при этом полезность представленной в анализе лингвистической информации. Прокомментируйте возможности применения подобных систем анализа.

Предложение, англ.	**Лингвистическая информация, представлена в синтаксическом анализе**
1.	
2.	
3.	
Общий комментарий	

4. Посетите сайт http://teneta.rinet.ru/hudlomer, помогающий определить функциональный стиль текста. Поместите в поле ввода любой отрывок текста объемом от 75 до 500 слов (примерно от 3 абзацев до 1 страницы)

   а) из вашей курсовой работы или реферата;

   б) из художественного произведения (используйте для этого, например, библиотеку М. Мошкова http://lib.ru);

   в) газетный текст (используйте текст любого сетевого СМИ, например, www.rg.ru).

   Оцените результаты автоматического определения стиля. Что вы думаете о возможностях такой системы?

5. Перейдите по ссылке http://www.antiplagiat.ru/QuickCheck.aspx и введите текст из вашей актуальной курсовой работы или реферата. Впишите ре-

зультат и ваш комментарий получившейся статистики в таблицу. Для каких целей можно использовать данную программу?

Результат:
Ваш комментарий:

Лабораторная работа 9

Автоматический синтез диалогов[1]

1. Побеседуйте на русском языке с виртуальным собеседником по адресу http://www.web4design.ru/virt_sobesednik.html. Постарайтесь узнать, сколько лет вашему собеседнику. Получили ли вы ответ? Оцените качество синтезируемых реплик и возможности использования данной программы.
2. Перейдите по ссылке http://www.beeline.ru/beelinebot/Default1.aspx и постарайтесь узнать у электронного помощника способы пополнения счета при нулевом и отрицательном балансе. Был ли этот помощник полезен для получения информации?
3. Побеседуйте на иностранном языке с виртуальным собеседником по адресу:

   английский: www-ai.ijs.si/eliza/eliza.html;

   немецкий: www.ego4u.de/de/chill-out/chat/egon-bot;

   французский: http://193.108.42.79/ikea-fr/cgi-bin/ikea-fr.cgi.

   Оцените дидактические возможности данной программы для обучения иностранному языку. Какой уровень знаний иностранного языка необходим для ее использования?

[1] Задания 1 и 2 составлены совместно с А.А. Кобелевым.

Часть 3

ПРИКЛАДНЫЕ РАЗДЕЛЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

3.1. Корпусная лингвистика

Корпусная лингвистика как раздел прикладной лингвистики. Понятие корпуса, разметки. Виды корпусов. Требования к корпусам.

Одной из важных задач лингвистики является сбор и хранение источников фактического материала для лингвистических исследований. 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 собраний текстов в электронном виде, или корпусов, является настолько значимой для современной лингвистики, что эти собрания электронных текстов становятся объектом исследований специального раздела прикладной лингвистики — корпусной лингвистики.

Корпусная лингвистика — раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов при помощи компьютеров [18, 3].

Исходя из такого определения можно констатировать, что 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 [47].

Современные исследователи-лингвисты могут как создавать свои собственные корпусы, а затем проводить необходимые иссле-

дования на их базе, так и использовать общедоступные корпусы, созданные другими исследователями и их коллективами.

Кроме проведения научных исследований корпусы могут использоваться [20, 166—167; 30, 60]:

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.
6) *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.

В более широком смысле корпусом считается любое собрание текстов. В этой трактовке выделяются размеченные (аннотированные) и неразмеченные корпусы текстов. 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, кото-

рая может оказаться нерелевантной для исследователя, что значительно затрудняет работу с таким источником. В связи с этим предметом корпусной лингвистики являются преимущественно размеченные корпусы 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 [4, 118—119, 47]:

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 корпусы 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);
6) *компьютерная поддержка* (поддержка корпуса текстов комплексом программ по обработке данных, обеспечивающих выявление контекстов слова, статистическую инвентаризацию, автоматическую словарную обработку и т.д.).

Важным этапом создания корпуса является его разметка. Разметка (англ. *tagging*, *annotation*) — это приписывание текстам и их компонентам специальных меток (англ. *tag*). Эти метки могут быть внешними (экстралингвистическими), включающими сведения об авторе и о тексте, или внутренними: структурными или собственно лингвистическими. *Внешние метки* содержат сведения об авторе, названии текста, годе и месте издания, жанре, тематике. Сведения

об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое. 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 [18, 6].

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 [18, 6—7].

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 [20, 170].

Для разметки корпуса сообщений Твиттера при проведении международного исследовательского проекта по изучению данного жанра (Ганновер, 2010) нами 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. 2).

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.

Таблица 2

Классификация корпусов [18, 13]

№	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.

Универсальный национальный корпус — это собрание текстов конкретного естественного языка, представительное по отношению ко 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) (www.ruscorpora.ru). Среди корпусов славянских языков выделяется Чешский национальный корпус (http://ucnk.ff.cuni.cz), созданный в Карловом уни-

верситете Праги. Национальные корпусы 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 [20, 169; 47].

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 [18, 3]. 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.

Вопросы для обсуждения на семинарских занятиях

1. Что может являться единицей корпуса?
2. Как отбираются тексты для корпуса? Проиллюстрируйте принципы отбора на примере Брауновского и других корпусов.
3. Дополните классификацию корпусов, представленную в пособии. Поясните, что означает «исследовательский корпус», «статический корпус», «параллельный корпус».

4. 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.ruscorpora.ru),
   - Национальный корпус русского литературного языка (www.narusco.ru),
   - Компьютерный корпус текстов русских газет конца XX века (www.philol.msu.ru/~lex/corpus)
   - Словарь-корпус языка А.С. Грибоедова (www.inforeg.ru/electron/concord/concord.htm)
   - Корпус института немецкого языка в Мангейме (www.ids-mannheim.de/kl/).
5. Составьте глоссарий по теме «Корпусная лингвистика». Используйте для этого рекомендуемые источники литературы и сетевые ресурсы. Включите в глоссарий определения следующих понятий: конкорданс, рандомизация, коллокация, подмассив, парсинг, лемматизация, корпус-менеджер.
6. Найдите сетевые ресурсы по теме «корпусная лингвистика» и кратко охарактеризуйте их.

Рекомендуемая литература

1. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. 3-е изд. М.: ЛКИ, 2007. С. 112—137.
2. Захаров В.П. Корпусная лингвистика: учебно-метод. пособие. СПб.: СПбГУ, 2005. С. 3—14.
3. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 166—173.
4. Овчинникова И.Г., Угланова И.А. Компьютерное моделирование вербальной коммуникации: учебно-метод. пособие. М.: Флинта: Наука, 2009. С. 60—76.

Лабораторная работа 10

1. Откройте веб-страницу Русского национального корпуса (РНК) (www.ruscorpora.ru), Корпуса русского литературного языка (КРЛЯ) (www.

narusco.ru) и Британского национального корпуса (БНК) (www.natcorp.ox.ac.uk). Введите в строку поиска этих корпусов слово *русский / Russian*. Заполните таблицу.

	РНК	**КРЛЯ**	**БНК**
Количество вхождений			

Как вы можете прокомментировать полученные результаты?

2. Выпишите 3 любых контекста использования слова *русский / Russian* в трех рассмотренных корпусах. Укажите источник каждого примера

№ примера	**РНК**	**КРЛЯ**	**БНК**
1			
2			
3			

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

№ примера	**РНК**	**КРЛЯ**	**БНК**
1			
2			
3			

4. 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.

№ примера	**РНК**	**КРЛЯ**	**БНК**
1			
2			
3			

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

6. Как можно использовать рассмотренные корпусы в лингвистическом исследовании?

3.2. Компьютерная лексикография

Понятие компьютерной лексикографии. Электронный словарь. Состав словарной статьи. Виды электронных словарей. Преимущества электронных словарей. Перспективы компьютерной лексикографии.

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

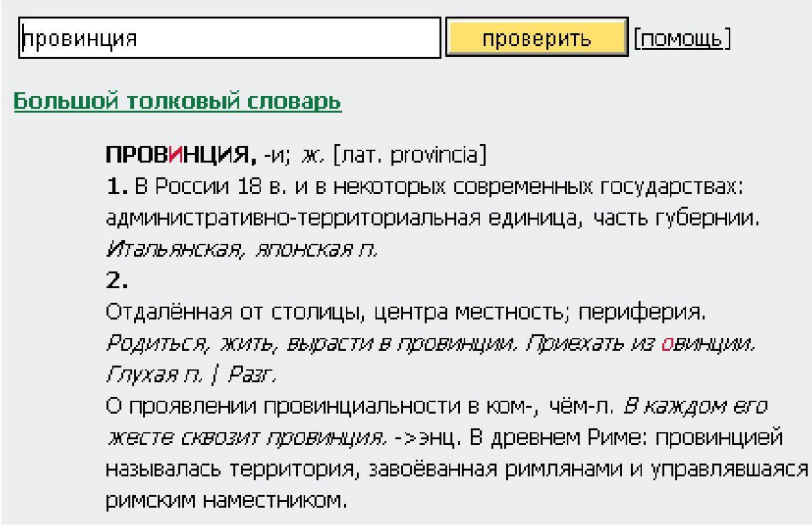
Основными задачами традиционной и компьютерной лексикографии являются определение структуры словаря и зон словарной статьи, а также разработка принципов составления различных видов словарей.

Словарь традиционно определяется как организованное собрание слов с комментариями, в которых описываются особенности структуры и/или функционирования этих слов [4, 55]. *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) [4, 86].

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. [4, 86]. Пример статьи словаря такого типа представлен на рис. 6.


провинция проверить [помощь]

Большой толковый словарь

ПРОВИНЦИЯ, -и; *ж.* [лат. provincia]
1. В России 18 в. и в некоторых современных государствах: административно-территориальная единица, часть губернии. *Итальянская, японская п.*
2.
Отдалённая от столицы, центра местность; периферия. *Родиться, жить, вырасти в провинции. Приехать из провинции. Глухая п. | Разг.*
О проявлении провинциальности в ком-, чём-л. *В каждом его жесте сквозит провинция.* ->энц. В древнем Риме: провинцией называлась территория, завоёванная римлянами и управлявшаяся римским наместником.

Рис. 6. **Статья компьютерного словаря [Большой толковый словарь: www.gramota.ru]**

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. [4, 86].

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 [4, 75—76].

В электронных словарях из названных компонентов обязательным является, пожалуй, лишь *словник*, в онлайн-словарях нередко имеется также *алфавит* с заложенными за каждой буквой 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.

Отличие электронных словарей от «бумажных» касается также их мультимедийности и гипертекстуальности: эти свойства выражены в электронных словарях в значительно большей степени, чем в печатных. Так, гиперссылки могут быть заложены за любым элементом словарной статьи или пунктом программного меню словаря. Это дает пользователю дополнительные возможности по поиску и быстрому переходу к необходимой словарной информации, позволяя найти синонимы и антонимы к заданному слову, слова той же семантической группы, парадигмы склонения и спряжения и т.д.

Гиперссылки позволяют также легко 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.

Отдельные электронные словари имеют также дополнительные возможности, например, электронный многоязычный словарь ABBYY

Lingvo x3 (© 2008 ABBYY) предоставляет *функцию обучения* (ABBYY Lingvo Tutor), позволяющую запоминать слова, отобранные по конкретной теме и представленные парами: русское и иностранное слово, составлять новые словари и словарные карточки, сохранять результаты обучения в файл и т.д.

В итоге структура электронного словаря в значительной степени отличается от структуры словаря печатного, хотя основная часть словаря — словник со словарными статьями — продолжает составлять ядро словаря в обоих случаях.

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.

Правда, 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.

Количество зон словарной статьи компьютерного словаря обычно превышает количество зон словарной статьи «бумажного» словаря, что обусловлено значительными ресурсами памяти и высокой скоростью обработки цифровой информации современными компьютерами. Но объем предлагаемой словарной информации должен соответствовать виду словаря: если читателю нужно произношение,

то «лишняя» информация о переводе проверяемого слова или его контекстных значениях будет только мешать пользователю.

Классификацию компьютерных словарей можно осуществлять на тех же принципах, что и классификацию обычных словарей. Традиционно выделяются лингвистические, энциклопедические и промежуточные (лингвострановедческие и терминологические) словари. В *лингвистических словарях* описываются сами слова — их значения, особенности употребления, структурные свойства, сочетаемость, соотношение с лексическими системами других языков и т.д. В *энциклопедических словарях* описываются понятия, факты и реалии окружающего мира, т.е. экстралингвистическая информация. Промежуточный тип словарей включает информацию и лингвистического, и экстралингвистического рода [4, 59—60].

Среди лингвистических словарей можно выделить несколько их видов [4, 59—74]:

- *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. [4, 65];
- *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» [39];

- *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 отмирании и видоизменении, 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 [5], 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 ABBYY 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 [4, 84]:

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.

Конечно, приведенное описание процесса создания электронного словаря может корректироваться в зависимости от его вида, исследовательских принципов и других факторов, ср. комментарии создателей электронного исторического словаря русского языка [48]. 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 под-

ходят 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 Page-Maker или WinWord, которые позволяют приписывать стили зонам словарных статей, алфавитизацию, создание указателей и т.д. [4, 82—85].

Пожалуй, единственный пример специализированной компьютерной программы, предназначенной для компьютерных лексикографических работ, является «Программа автоматизированного составления и обработки словников» (авторы: М.В. Литус, Е.В. Литус). Эта программа достаточно активно используется в филологических исследованиях и подробно представлена в учебном пособии А.Т. Хроленко и А.В. Денисова [52, 52—63].

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 [40]:

1) электронные словари позволяют по-разному представить содержание словарной статьи (различные «проекции» словаря), в том числе с помощью разнообразных графических и мультимедийных средств, которые не используются в обычных словарях;

2) в выдаваемой информации находят отражение различные технологии компьютерной лингвистики, например морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п.;

3) становится возможным быстро получить информацию, которая содержится где-то в недрах словаря и непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме;

4) электронный словарь позволяет быстро реагировать на изменения в языке и мире, и выпуск каждой последующей его версии или внесение изменений в онлайн-версию не занимает много времени и труда.

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» и т.п.
- Не нашла отражение в массовой лексикографической практике проблема *описания семантики* и практической *реализации грамматического словоизменения* и *словообразования*. Каждый язык имеет свои собственные способы грамматического кодирования смысла, которые не описываются в массовых словарях систематически. Например, как передать по-английски смысл «довыпендриваться», даже если знаешь, как передать «выпендриваться»?
- В словарях не существует даже системы понятий, с помощью которой *синтаксическая информация* могла бы быть доведена до обычного читателя. Решением этой проблемы могли бы стать интегральные словарные описания, основанные на формальных моделях, учитывающие прогрессивные лексикографические идеи. На этих же моделях следует организовать технологии доступа к словарному содержанию [40].

Названные проблемы могут быть решены при сотрудничестве лексикографов-теоретиков и практиков, а компьютерные инструменты, несомненно, облегчат рутинную работу по осуществлению монотонных лексикографических операций.

В целом констатируем, что компьютерная лексикография, направленная на создание электронных словарей, представляет собой весьма перспективное и нужное направление компьютерной лингвистики, поскольку создаваемые ею продукты — электронные словари — отличаются многогранностью, мультимедийностью, интеграцией новейших технологических решений, актуальностью материала и отвечают потребностям пользователя в организации доступа к необходимой информации.

Вопросы для обсуждения

1. Представьте структуру машинной словарной статьи.
2. Опишите зону морфологических сведений. Какие кодировки используются для обозначения частей речи и представления морфологической информации?
3. Чем различаются зона семантических и зона лексических сведений машинной словарной статьи? Проиллюстрируйте различия примерами.
4. Дайте определение базы данных. Что такое «данные»? Каковы основные способы организации баз данных?
5. Опишите особенности электронных переводческих словарей *ABBYY Lingvo* и *Multitran*. Чем они отличаются от онлайн-переводчиков (Google, Yandex и т.п.)?

Рекомендуемая литература

1. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. М.: Эдиториал УРСС, 2007. С. 55—87.
2. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 82—98, 146—153, 158—163.
3. Селегей В.П. Электронные словари и компьютерная лексикография // Ассоциация лексикографов Lingvo www.lingvoda.ru/transforum/articles/selegey_a1.asp
4. Егоров А. Слово за слово // Домашний компьютер. 2007. № 9. http://offline.homepc.ru/2007/135/334406

Лабораторная работа 11

Электронные словари

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.

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.

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.

Зоны словарной статьи	**Данные для слова *культура* 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?

5. Сравните количество зон словарной статьи в электронной и бумажной версиях словаря Ожегова. В какой версии представлено больше зон словарной статьи? Какие нужные, на ваш взгляд, зоны словарной статьи отсутствуют в той или другой версии? С каким словарем вам было удобнее работать и почему?

3.3. Компьютерная терминография

Понятие компьютерной терминографии. Термин как основной объект терминографии. Терминологические банки 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 те 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 [4, 89]. 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. В связи с этим задача унификации терминов и терминосистем должна быть обязательной составляющей государственной и местной языковой политики, поскольку многозначность и омонимия терминов, отсутствие согласования между близкими терминосистемами, создание терминологических сочетаний с труднопроизносимыми и неблагозвучными аббревиатурами (ср. *ГИБДД*) являются ощутимым препятствием для научно-технического прогресса [4, 90].

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 [4, 91—104].

Большинство электронных терминологических словарей носит дескриптивный характер и представляет термины отдельной отрасли знания. При этом востребованы и толковые (одноязычные), и переводные (двуязычные или многоязычные) словари. Разнообразные терминологические словари русского языка (анатомический, экономический, 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 [4, 105—106; 41, 122]:

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 нетерминами;
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.

Так, по своему происхождению термины могут быть заимствованными; в этом случае они переводятся, ср.: нем. Leitung → управ-

ление, или транслитерируются, например: англ. *Computer* → *компьютер*. Кроме того, термины могут образовываться из словообразовательных элементов родного языка (приставка) или путем изменения семантики существующих слов (*поле* → *(семантическое) поле*).

При анализе составляющих словарной статьи терминологического словаря можно заметить, что такой словарь требует еще более тщательной работы, чем обычный словарь.

Вопросы для обсуждения

1. Охарактеризуйте терминографическую традицию разных стран. Какие выводы можно сделать из этого сравнения?
2. В чем заключаются требования к специальным словарям?
3. Дайте определения известным вам видам терминологических словарей. Чем отличаются дескриптивные и нормативные терминологические словари?
4. Что входит в зоны словарной статьи терминологического словаря? Опишите одну из таких зон подробнее.

Рекомендуемая литература

1. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. 3-е изд. М.: ЛКИ, 2007. С. 90—95.
2. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 163—166.
3. Марчук Ю.Н. Компьютерная лингвистика: учеб. пособие. М.: АСТ: Восток — Запад, 2007. С. 190—195.

Лабораторная работа 12

Компьютерная терминография

1. Откройте главную страницу Европейского интерактивного терминологического банка данных IATE (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.

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

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

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

Прилагательные	Существительные
информационный мультимедийный цифровой электронный	ресурс технология средства платформа

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

Термин	Словосочетания с данным термином
ресурс	
технология	
средства	
платформа	

6. 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.

Зоны словарной статьи	Описание
Лексический вход	
Определение	
Примеры	
Источники	

3.4. Машинный перевод

Понятие перевода и машинного перевода. Классификация систем МП. Системы переводческой памяти. Этапы осуществления полностью автоматизированного МП. Проблемы МП. Примеры систем МП. Параметры оценки систем МП.

Вопросы машинного перевода составляют одну из центральных областей использования информационных технологий в лингвистике. Это обусловлено не только тем, что в машинном переводе как в фокусе концентрируются все проблемы компьютерной лингвистики — от способов анализа содержания до синтеза словоформы, предложения и целого текста [25, 14], но и постоянно возрастающей практической потребностью современного общества в переводе значительного количества текстов различной функциональной направленности.

Так, свыше 5 млрд жителей Земли используют около трех тысяч языков, и все большее их количество включается в мировые информационные потоки. Разноязычная информация, накапливаемая в геометрической прогрессии, становится труднодоступной, так как на поиск и перевод нужных сведений требуются значительные материальные затраты. Было подсчитано, что если синтез нового химического соединения обойдется менее чем в 100 тыс. долларов, выгоднее произвести этот синтез, чем искать описание аналогичной работы на других языках [52, 114].

Другой иллюстрацией возрастания потребности в переводе служат документы международных организаций, которые в обязатель-

ном порядке переводятся на языки стран-участников. Только Европейский союз в настоящее время объединяет 27 государств, в которых используется 23 официальных языка (http://europa.eu). Это обеспечивает работой несколько тысяч профессиональных переводчиков, переводящих в год миллионы страниц. Услуги переводчиков обходятся в миллиарды долларов.

Кроме того, что работа переводчика-человека достаточно дорогая, она к тому же весьма медленная. Так, нормой научно-технического перевода считается время 10 дней на авторский лист (24 страницы машинописного текста) [26, 4]. Система машинного перевода позволяет получить перевод сотен авторских листов за 1 час [20, 79].

Кроме того, появляются новые области применения машинного перевода, например, тексты Интернета. По подсчетам исследователей, в Интернете встроенными системами перевода (SYSTRAN, TRADOS и ESTeam Translator) и сетевыми онлайновыми словарями ежедневно выполняется 1 млн запросов на перевод текстов в различных форматах [8, 102].

Все вышесказанное свидетельствует об актуальности обращения к проблеме машинного перевода, который хотя и уступает по качеству переводу, осуществляемому человеком, но даже на сегодняшнем этапе развития позволяет преодолевать языковые барьеры, а кроме того, продолжает оставаться интересной научной проблемой компьютерной лингвистики в целом.

Чтобы определить понятие машинного перевода, обратимся сначала к некоторым общим понятиям теории перевода. *Перевод* обычно понимается как деятельность, «в результате которой некоторый текст на одном языке ставится в соответствие тексту на другом языке, при этом обеспечивается их смысловая эквивалентность» [23, 30]. При этом отмечается многозначность понятия перевода: это одновременно и процесс передачи содержания текста на одном языке средствами другого языка, и результат переводческой деятельности [4, 138].

Перевод представляет собой весьма сложный вид интеллектуальной деятельности человека, поскольку это не чисто языковой, а сложный когнитивный феномен: в процессе перевода человек использует лингвистические и экстралингвистические знания, а кроме

того, в этот процесс включаются два принципиально различных этапа: понимание текста на исходном языке (ИЯ) и синтез текста на языке перевода (ПЯ) [4, 138].

Вследствие такой комплексности переводческого процесса наука о нем (*переводоведение*) носит междисциплинарный характер и оказывается связанной с лингвистикой, литературоведением, когнитивными науками и культурной антропологией [4, 138]. В частности, исследователями отмечается, что переводятся не столько слова и их последовательности, сколько мыслительные образы, порождаемые в сознании переводчика под их воздействием [7, 152], т.е. связь перевода и когнитивных, мыслительных процессов человека очевидна.

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 [24, 15].

С точки зрения роли человека в процессе выполнения МП различают следующие его виды [15, 54]:

- MAHT (*Machine-assisted human translation*) — перевод, осуществляемый человеком с использованием компьютера;
- HAMT (*Human-assisted machine translation*) — машинный перевод при участии человека;
- FAMT (*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*, *TM*). 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 куска 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.

Третий тип систем МП является наиболее сложным, поэтому остановимся на его характеристике подробнее.

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 [7, 152; 20, 79], 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 [6, 93]:

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 (поверх-

ностный, 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 [6, 93—94].

В итоге в целом схема машинного перевода включает следующие этапы [20, 80—81; 6, 94]:

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*) [24, 15], снятия омонимии на разных уровнях, а также с необходимостью привлечения в процесс перевода экстралингвистических знаний [8, 116, 119].

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 [12].

О снятии омонимии говорилось ранее, необходимость же включения экстралингвистической информации в процесс МП иллюстрируется, к примеру, следующими фразами [цит. по: 8, 120]:

Председатель Центральной избирательной комиссии ***назначается*** *президентом Российской Федерации.*

Согласно задумкам американских ученых, сразу после старта вражеские ракеты ***будут уничтожать*** *авиационные лазеры и мобильные комплексы малых противоракет.*

Лишь знания о соответствующих предметных областях позволяют в данном случае определить типы глубинных синтаксических отношений *председатель — президент* ('председатель становится президентом' или 'президент назначает председателя') и лазеры — ракеты ('лазеры уничтожают ракеты' или наоборот).

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 [6, 94—95].

К числу наиболее распространенных в России систем МП относятся [6, 95]:

- *Stylus* — система МП, включающая множество словарей по разным предметным областям;
- *Universal Translator* — многоязычная система МП;
- *Socrat* — система, позволяющая сканировать документы, переводить их содержимое и проверять орфографию;
- *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, FEMTI*) [8, 106—107]:

- 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. [6, 91]. Соответственно, и оценку каждой из систем МП нужно производить с учетом их названных особенностей.

Итак, машинный перевод, представляющий собой процесс передачи содержания текста на одном языке средствами другого языка с использованием компьютеров, является одним из первых и не теряющих своей актуальности направлений компьютерной лингвистики. Процесс машинного перевода может предполагать разную степень активности человека в его выполнении, что обусловливает многообразие его форм, выбор которых зависит от целей перевода и его условий.

Вопросы для обсуждения

1. Исследователи считают, что причины появления и развития идеи МП лежат в технической, политической и социальной областях. Поясните каждую из причин.
2. Как вы можете объяснить связь процесса машинного перевода и дешифровки текстов?
3. Охарактеризуйте этапы развития МП. Какую роль в развитии идеи МП сыграл американский ученый У. Уивер?
4. Какую роль человек может играть в процессе машинного перевода? Что такое предредактирование и постредактирование?
5. В чем, на ваш взгляд, заключается будущее МП?

Рекомендуемая литература

1. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. 3-е изд. М.: ЛКИ, 2007. С. 168—173.
2. Беляева Л.Н. Лингвистические автоматы в современных гуманитарных технологиях: учеб. пособие. СПб.: Книжный Дом, 2007. С. 102—132.
3. Всеволодова А.В. Компьютерная обработка лингвистических данных: учеб. пособие. 2-е изд., испр. М.: Флинта: Наука, 2007. С. 53—63.
4. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 75—81, 108—109.
5. Овчинникова И.Г., Угланова И.А. Компьютерное моделирование вербальной коммуникации: учебно-метод. пособие. М.: Флинта: Наука, 2009. С. 80—91.

Лабораторная работа 13

Машинный перевод

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.

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

2. Охарактеризуйте протестированные онлайн-переводчики по следующим параметрам:

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, характеризующую редактирование. При необходимости дополните таблицу собственными параметрами.

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)

Тема	Комментарии
1. Компьютеры	
2. Бизнес	
3. Здоровье	

3.5. Компьютерное обучение языкам

Информатизация образования и связанные с этим изменения в обучении языкам. Понятие компьютерного обучения языкам. Классификация электронных средств обучения. Дистанционное обучение. Сетевые ресурсы в обучении языкам.

Одной из важных практических областей применения компьютеров в лингвистике является компьютерное обучение языкам (Computer Assisted Language Learning, CALL). Компьютеризация и информатизация являются характерными особенностями современного обучения в целом, поскольку применение современных информационно-коммуникационных технологий в обучении позволяет сделать его более эффективным, повысить мотивацию обучающихся и сократить затраты человеческого труда. Кроме того, применение компьютеров в полной мере соответствует другим современным тенденциям образования: его деятельностному и личностно-ориентированному характеру.

При этом обучению иностранным языкам (ИЯ) с помощью компьютера отводится особая роль, так как изучение языка представляет собой процесс, использующий весь спектр человеческих возможностей познания [34, 38].

Вопросы использования компьютеров в обучении рассматриваются с 1950-х годов, т.е. практически с начала промышленного производства компьютеров [38, 3]. За более чем полувековой период компьютерное обучение претерпело значительные изменения, которые определялись господствующим методом обучения и уровнем развития компьютерной техники. Так в развитии компьютерных обучающих средств выделяют два главных этапа:

1) бихевиористический: на этом этапе обучающие программы были построены по формуле «стимул — реакция», обучающемуся отводилась пассивная роль объекта обучения, а программы выполняли функцию тренажеров;

2) когнитивно-интеллектуальный: программы ориентированы на обучающегося, дают ему свободу выбора уровня и типа действий, активизируя тем самым его познавательные функции [20, 111; 34, 38—39].

В настоящее время компьютерное обучение ИЯ представляет собой отдельную область знаний и практических действий, нацеленных на использование компьютеров в обучении и изучении языков [56, 1], имеющую свою методику, программные средства, цели и задачи. Возможности использования компьютеров простираются от традиционных программ-тренажеров до современных виртуальных обучающих сред, мультимедийных программ и применения различных форм общения и хранения информации в Интернете, в частности электронной почты, корпусов и конкордансов, подкастов и т.п., с дидактическими целями.

Особенностью компьютерного обучения языкам является то, что это обучение опирается на определенный *теоретический метод* (бихевиористский, коммуникативный, когнитивно-интеллекутальный и т.п.), а кроме того, носит *междисциплинарный* характер: в наши дни проблемы компьютерного обучения языкам решаются совместными усилиями психологов, методистов, программистов, вебдизайнеров и лингвистов. В таком междисциплинарном сотрудничестве возникают новые оригинальные подходы к компьютерному обучению. В частности, заимствование идей и методов из сферы искусственного интеллекта породило новое направление в компьютерном обучении языкам — ICALL (*Intelligent Computer Assisted Language Learning*) [34, 39; 58].

Компьютеры могут использоваться в обучении языкам различным образом:

1) *компьютер — помощник преподавателя* (использование компьютера преподавателем на отдельных этапах традиционного занятия);
2) *компьютер — преподаватель* (индивидуальное обучение целому учебному курсу по заданному жесткому сценарию);
3) *компьютер — источник и «оцениватель» знаний* (групповое и индивидуальное обучение в рамках дистанционного обучения языкам, при котором обучающийся сам обращается к компьютеру как к носителю необходимой информации и «оценивателю» приобретенных знаний) [20, 138—139].

Очевидно, что у каждого способа компьютерного обучения есть свои целевые группы и условия: обязательные для изучения курсы тре-

буют привлечения компьютера в помощь преподавателю, а удовлетворение индивидуальных потребностей в образовании — использования компьютеров в качестве преподавателей или источников и «оценивателей» знаний.

Во всех случаях используются разнообразные электронные обучающие средства [34, 18—31; 60, 190—196; 61]:

- компьютерные учебники;
- тестирующие программы;
- тренажерные программы;
- учебные игры;
- компьютерные справочники и энциклопедии и др.

Компьютерный учебник — это программно-методический комплекс, позволяющий самостоятельно освоить учебный курс или его большой раздел. Он объединяет в себе свойства обычного учебника, справочника, задачника и лабораторного практикума и представляет собой не альтернативу, а дополнение к традиционным формам обучения.

Тестирующая программа — это компьютерная программа, предлагающая пользователю вопрос и несколько вариантов ответов на него. Основная задача такой программы — проверка знаний пользователя. Наиболее простые тесты имеют фиксированное количество стандартных вопросов и неизменную систему оценки полученных ответов.

Тренажерная программа — программа формирования автоматического навыка выполнения определенных коммуникативных действий путем многочисленных повторов таких действий. Примерами программ такого типа служит обучение быстрому набору текста на клавиатуре методом слепой печати, упражнения на употребление правильных форм глагола и т.д.

Учебные игры — это компьютерные программы, имеющие игровые и обучающие функции. В этом случае учащиеся активно вовлекаются не только в процесс получения, но и использования знаний, выполняя какую-либо фиктивную роль. Например, создана учебная игра по интерактивному освоению ландтага немецкой федеральной земли

Нижняя Саксония. Каждый участник такой игры становится виртуальным экскурсантом по ландтагу и после получения информации об определенной части этого законодательного органа, успешно выполнив тестовые задания, может переходить на другой этаж здания.

Учебные игры можно применять с различными целями:

- для мотивации учащихся в получении новых знаний,
- для отдыха, развлечения, снятия напряжения на уроке,
- для активизации интереса учащихся,
- для активизации познавательной самостоятельности,
- для отработки умений учащихся, как тренажер.

Компьютерные справочники и энциклопедии — программы, предназначенные только для представления учебного материала. Обычно они содержат очень большой объем информации, что требует обязательного использования автоматического поиска. Еще одной отличительной особенностью данного вида электронных обучающих ресурсов является мультимедийный характер представляемой в них информации и ее гипертекстовая организация.

Обучающие компьютерные ресурсы могут предлагаться уже готовыми или создаваться самими преподавателями с помощью заготовок несложных компьютерных упражнений. Примерами ресурсов первого типа выступают, в частности, следующие мультимедийные обучающие программы:

- *Профессор Хиггинс*. Английский без акцента: мультимедийное учебное пособие по английской фонетике и грамматике (www.istrasoft.ru/higgins/htm);
- *Bridge to English*: программа по обучению английской лексике и грамматике для взрослых (www.intense.ru).

Программной оболочкой, позволяющей составлять несложные упражнения в виде кроссвордов, предложений с пропущенными словами, текстов с перемешанными предложениями и т.д. является, к примеру, программа *Hot Potatoes* (http://hotpot.uvic.ca).

Электронные обучающие ресурсы разного рода составляют основу современного дистанционного обучения. *Дистанционное обу-*

чение — это форма организации учебного процесса, основывающаяся на принципе самостоятельного получения знаний, предполагающая телекоммуникационный принцип доставки учебного материала и интерактивное взаимодействие обучающихся и преподавателей в процессе обучения и при оценке знаний [20, 142—143].

Дистанционное обучение обычно предполагает регистрацию (запись на курс), позволяющую организовать обратную связь с обучающимся, предоставление обучающемуся учебных материалов разного рода (текст, иллюстрации, видео, задания и т.д.) и выполнение тестовых заданий, позволяющих оценить уровень знаний обучающегося. Нередко обучающийся может получать консультации преподавателя курса в чате или по электронной почте.

При организации дистанционного курса особую роль играет его рациональное построение: выделение отдельных тем, отбор теоретического материала, заданий и упражнений для каждой темы, гибкая система тестовых заданий.

Огромным по важности разделом современного компьютерного обучения языкам становится использование различных веб-ресурсов (электронных писем, веблогов, подкастов, совместных вики-проектов и т.п.) с дидактическими целями. В целом *веб-ресурс* можно понимать как электронный документ, содержащий информацию различного рода (вербальную, графическую, табличную, звуковую, графическую, видеофайлы, анимацию и компьютерные программы), доступную через веб-страницы, размещенные во Всемирной паутине [38, 340].

Несомненно, что такие свойства веб-ресурсов, как доступность, обширность, глобальность и аутентичность делают их удобным источником учебного материала для обучения языкам. В то же время очевидно, что веб-ресурсы могут быть весьма разными, а их количество во Всемирной паутине растет в геометрической прогрессии. В этих условиях при желании использовать веб-ресурсы с целью обучения языкам очень важным становятся принципы их отбора и задания для обучающихся, направленные на поиск необходимой учебной информации во Всемирной паутине. Примерами заданий такого типа являются, в частности, *веб-квесты*, понимаемые как сценарии организации проектной деятельности учащихся по любой теме с ис-

пользованием сети Интернет [44, 97]. Темы веб-квестов, используемых при обучении иностранному языку, могут быть самыми разными: *My career* [14, 98—101], *Extreme Sports*[1], *Джордж Гордон Байрон*[2].

Веб-квесты относятся к такому типу веб-ресурсов, который получил название Веб 2.0 (социальная сеть). К Веб 2.0 относятся социальные сервисы и службы Всемирной паутины, позволяющие широкому кругу людей быть не только получателями информации, но и ее создателями и соавторами [44, 26—31].

В заключение раздела констатируем, что компьютерное обучение языкам — это весьма перспективное направление современной лингводидактики. При этом не следует рассматривать компьютерные обучающие ресурсы как замену преподавателя, а считать их способом расширения традиционного занятия для организации и выполнения рутинной работы, развития навыков обучающихся путем тренировки, повышения активности обучающихся и создания возможностей для самообразования.

Вопросы для обсуждения

1. Охарактеризуйте бихевиористский и когнитивно-интеллектуальный подходы в компьютерном обучении языкам.
2. В чем заключаются преимущества и недостатки использования компьютерных обучающих ресурсов?
3. Опишите этапы создания мультимедийных обучающих программ.
4. Назовите параметры классификации мультимедийных обучающих программ.
5. В чем заключаются преимущества и недостатки дистанционного обучения?
6. Кратко охарактеризуйте следующие виды веб-ресурсов: образовательные порталы, электронные библиотеки, журналы в электронной версии.

Рекомендуемая литература

1. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 110—114, 138—145.

[1] www.longwood.k12.ny.us/wmi/wq/werner2/index.htm

[2] www.spodon.ru/quest/biron/index.html

2. Потапова Р.К. Новые информационные технологии и лингвистика. Изд. 2. М.: Эдиториал УРСС, 2004. С. 18—42, 117—118.

3. Хроленко А.Т., Денисов А.В. Современные информационные технологии для гуманитария: практ. руководство. М.: Флинта: Наука, 2007. С. 10—30, 64—100.

Лабораторная работа 14

Компьютерное обучение языкам[1]

1. Перейдите по ссылкам:
   - www.cambridge.org/us/esl/students/?site_locale=en_US#
   - http://www.arcademicskillbuilders.com (Language Arts)

   Ознакомьтесь с приложениями для изучения английского языка *Connect Arcade, Skychase, Furious Frogs, Spelling Bees*. Определите вид мультимедийной программы и теоретический подход, использованный при ее создании. Заполните таблицу.

Программа	Вид	Подход	Обоснование
Connect Arcade			
Skychase			
Furious Frogs			
Verb Viper			

2. Перейдите на сайт Интернет-Университета Информационных Технологий по ссылке www.intuit.ru. Зарегистрируйтесь на сайте, выберите один из бесплатных дистанционных курсов и запишитесь на него. Изучив информацию о курсе, заполните таблицу.

Название курса	Автор курса	Цель	Уровень	Кол-во часов

3. Просмотрите выборочно несколько модулей курса. Заполните таблицу.

Параметр	Описание
Составляющие курса (вид: урок, тема или др., количество, примеры)	
Возможные трудности	
Подтверждение (сертификат, свидетельство, диплом)	

[1] Задания лабораторной работы разработаны совместно с А.А. Кобелевым.

4. Какие из перечисленных веб-ресурсов не являются порталами:

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

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

Вид ресурса	Русский язык	Иностр. язык
Электронная библиотека		
Электронный журнал		

3.6. Информационно-поисковые системы

Понятие информационно-поисковой системы. Виды поисковых средств в Интернете. Характеристика поисковой системы Интернета. Информационно-поисковый язык.

В современном мире, который буквально пронизан постоянно нарастающими объемами информации, для человека, использующего эту информацию с целью ее превращения в знания, встает проблема ориентации. Чтобы не захлебнуться в информационном потоке, нам нужны техники отбора, фильтрации и оценки [10, 18].

Традиционными способами фильтрации и отбора 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 [6, 70].

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 [4, 197].

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) [19, 3].

Для экономии усилий человека с 1950-х годов осуществляются попытки создания автоматизированных ИПС. При этом в первых ИПС анализ и описание содержания документов (индексирование) выполнялись вручную, а поиски по этим документам проводились автоматически [19, 8].

Сегодня с развитием компьютерной техники и созданием высокоскоростных телекоммуникационных сетей в деле автоматизации поиска достигнуты значительные успехи, кратко и емко выразившиеся в знаменитой формуле Б. Гейтса «информация на кончиках пальцев» (information at your fingertips) [цит. по: 10, 16]. Данное выражение можно понимать таким образом: информация всегда находится в распоряжении человека, нужно лишь сделать несколько нажатий клавиш, чтобы получить доступ к ней.

Так, для поиска информации в Интернете служат различные классы поисковых средств [6, 71]:

- каталоги (*directories*);
- подборки ссылок (*bookmarks*);
- поисковые машины (*search engines*);
- базы данных адресов электронной почты и т.д.

Каждый вид поискового средства имеет свои особенности, так, если человек имеет недостаточно точное представление о цели поиска, ему целесообразнее использовать каталоги веб-ресурсов. Применение поисковых машин эффективно, если пользователь представляет, какие ключевые слова характеризуют нужные ему ресурсы.

Каталог веб-ресурсов — это постоянно обновляемая и пополняемая система ссылок на ресурсы, распределенные по иерархической структуре категорий. На верхнем уровне каталога представлены самые общие категории (рубрики), например «наука», «бизнес», «развлечения» и т.д. На нижележащих уровнях рубрики имеют более частный характер [6, 71]. Например, рубрика «наука» может делиться на категории «точные науки», «естественные науки» и «гуманитарные науки», последние — на философию, социологию, психологию, педагогику и т.д. Русскоязычный каталог сайтов можно найти, например, по адресу www.ru.

Коллекция ссылок представляет собой еще один способ организации информации во Всемирной паутине. Такая коллекция обычно составляется специалистом в определенной теме, постоянно обновляется и не содержит ненужной информации. Печатный аналог такой коллекции ссылок по использованию информационных технологий в лингвистике можно найти после библиографического списка в нашем пособии. Некоторые примеры коллекций ссылок по обучению английскому языку приводит С.В. Титова [45, 27—28].

Поисковые машины (или поисковые системы) — это специальные веб-страницы, позволяющие находить веб-ресурсы, текстовое содержание которых соответствует запросу пользователя. В Международном каталоге поисковых машин (www.searchenginecolossus.com) зарегистрировано свыше 2300 поисковых систем из 232 стран. По данным этого каталога, каждый день выполняется до 450 млн поисковых запросов [6, 72; 38, 364].

К наиболее известным поисковым машинам относятся [6, 72—73]:

- AltaVista (www.altavista.com);
- Excite (www.excite.com);
- Yahoo! (www.yahoo.com);
- AOL (http://search.aol.com);

- MSN (http://search.msn.com);
- Google (www.google.ru);
- Яndex (www.yandex.ru);
- Rambler (www.rambler.ru);
- Апорт (www.aport.ru).

Рассмотрим, как осуществляется поиск в поисковой системе. Пользователь вводит свой поисковый запрос в специальную строку. 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 [4, 201]. Информационно-поисковые языки представляют собой знаковые системы со своим алфавитом, лексикой, грамматикой и правилами пользования. О специфике ИПЯ каждой поисковой системы, особенно о его «синтаксисе» (т.е. о правилах сочетания ключевых слов, вводимых в строку поиска) можно узнать на отдельных вкладках соответствующей поисковой системы. Например, в Яндекс такая вкладка называется «Помощь — Как искать».

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 поиска основаны на сравнении поискового предписания с поисковым образом запроса [4, 201].

Степень соответствия документа запросу задается категорией *релевантности*. При этом в процессе информационного поиска можно получить в выдаче значительный *информационный шум* — множество документов, формально релевантных, но не являющихся релевантными по смыслу [4, 197—198].

Чтобы получить меньше информационного шума, пользователю следует уточнять свой запрос, используя для этого дополнительные настройки поисковой системы. Так, в *Google*, нажав вкладку «Расширенный поиск», можно задать поиск целых словосочетаний (а не отдельных составляющих их слов), ограничить язык выдачи, дату

создания документа, часть документа, в которой используется слово, формат документа и т.д. Такие манипуляции увеличивают вероятность нахождения нужной информации уже в самом начале выдаваемого списка.

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 [19, 8].

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.

Итак, информация не просто дается человеку «на кончиках пальцев», а предполагает сложные и трудоемкие процессы сортировки и отбора. С этими задачами в значительной степени помогают справиться современные автоматические информационно-поисковые системы, в частности поисковые системы Всемирной паутины.

Вопросы для обсуждения

1. Что такое формальная и смысловая релевантность поиска? Как различие этих понятий отражается на результатах поиска?
2. Как вы понимаете пертинентность? Какие способы снижения пертинентности вы можете предложить?
3. Охарактеризуйте два основных типа информационно-поисковых систем: документальные и фактографические.
4. В чем состоят различия информационно-поисковых систем с ручным и автоматическим индексированием? Приведите примеры систем обоих типов.
5. Что такое общий и специализированный каталог веб-ресурсов? Приведите примеры каталогов обоих типов.
6. Что такое фасетная классификация? Приведите примеры фасетов при описании одного документа.

Рекомендуемая литература

1. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. М.: Эдиториал УРСС, 2001. С. 197—207.
2. Захаров В.П. Информационно-поисковые системы: учеб.-метод. пособие. СПб., 2005. С. 3—18.
3. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004. С. 183—185.

Лабораторная работа 15

Информационный поиск в Интернете

1. Ознакомьтесь с информационно-поисковым языком двух поисковых систем: *Google* и *Рамблер*, которые вы можете найти по ссылкам www.google.ru/intl/ru/help/refinesearch.html и http://help.rambler.ru/project.html?s=search
2. 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.

Параметр	**Google**	**Рамблер**
Запрос		
Документ, отвечающий результатам запроса (url)		
Номер этого документа в списке результатов		
Инф. шум (количество нерелевантных ссылок)		
Полнота (в Рунете всего 4 источника)		
Точность		
Выводы (результаты какой ИПС были более полными и точными, где было меньше информационного шума, синтаксис какой ИПС более комплексный, простой, удобный):		

3. Изучите информацию по использованию языка запросов в Яндексе (www.yandex.ru/info/syntax.html). Примените полученную информацию, приняв участие в Яндекс-Кубке (kubok.yandex.ru). Внесите результаты своей поисковой деятельности в таблицу.

Вопрос	**Время, потраченное на выполнение поиска**	**Ответ (url)**

ЗАКЛЮЧЕНИЕ

В соответствии с целью нашего пособия — дать краткий обзор основным возможностям использования информационных технологий в лингвистике — мы не можем представить все многообразие этих возможностей, выделяемых в компьютерной лингвистике, компьютерной лингводидактике и других областях. Наряду с рассмотренными здесь способами использования компьютеров (автоматический анализ и синтез устной речи, автоматический ввод текста, автоматический анализ текста, использование корпусов текстов, компьютерное обучение языкам и т.д.) существуют и другие области пересечения лингвистики и информатики: извлечение знаний из текста, автоматическое индексирование и рубрицирование документов, гипертекстовые технологии в лингвистике и многое другое.

Кроме того, можно расширять и углублять каждый раздел, поскольку по каждой теме опубликовано достаточно значительное количество научных и учебных работ, предлагаются веб-ресурсы и программные разработки.

Надеемся, что это пособие послужит первой ступенькой в освоении сложных, но интересных и перспективных вопросов использования информационных технологий в лингвистике и ее прикладных областях, и за вводным учебным курсом «Информационные технологии в лингвистике» последуют курсы специализации по автоматическому анализу текста, машинному переводу или компьютерному обучению языкам.

БИБЛИОГРАФИЯ

Список использованной научной литературы

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011.
2. Алексеев В. Услышь меня, машина // Компьютерра. 1997. № 49. http://offline.computerra.ru/1997/226/938 (дата обращения: 28.02.2012).
3. Апресян Ю.Д. Идеи и методы современной структурной лингвистики. М.: Просвещение, 1966.
4. Баранов А.Н. Введение в прикладную лингвистику: учеб. пособие. 3-е изд. М.: ЛКИ, 2007.
5. Баранов А.Н., Караулов Ю.Н. Русская политическая метафора: материалы к словарю. М.: ИРЯ, 1991.
6. Башмаков И.А, Башмаков А.И. Интеллектуальные информационные системы. М.: МГТУ им. Н.Э. Баумана, 2005.
7. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. М.: Русский мир, 2004.
8. Беляева Л.Н. Лингвистические автоматы в современных гуманитарных технологиях: учеб. пособие. СПб.: Книжный Дом, 2007.
9. Березин Ф.М., Головин Б.Н. Общее языкознание. М.: Просвещение, 1979.
10. Больц Н. Азбука медиа / пер. с нем. Л. Ионина, А. Черных. М.: Европа, 2011.
11. Большакова Е.И. Компьютерная лингвистика: методы, ресурсы, приложения // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. С. 90—105.
12. Бонч-Осмоловская А.А., Толдова С.Ю. Разрешение анафоры // Фонд знаний «Ломоносов». М., 2011. www.lomonosov-fund.ru/enc/ru/encyclopedia:0127469:article (дата обращения: 28.02.2012).
13. БЭС — Большой энциклопедический словарь. Языкознание. М.: Большая Российская энциклопедия, 1998.
14. Воробьева Е.И. Информатизация иноязычного образования: основные направления и перспективы. Архангельск: Поморский университет, 2011.

15. Всеволодова А.В. Компьютерная обработка лингвистических данных: учеб. пособие для студ., аспир., препод.-филологов. 2-е изд., испр. М.: Флинта: Наука, 2007.

16. Гейн А.Г., Сенокосов А.И. Справочник по информатике для школьников. Екатеринбург: У-Фактория, 2003.

17. Егоров А. Слово за слово // Домашний компьютер. 2007. № 9. http://offline.homepc.ru/2007/135/334406 (дата обращения: 28.02.2012).

18. Захаров В.П. Корпусная лингвистика: учеб.-метод. пособие. СПб.: СПбГУ, 2005а.

19. Захаров В.П. Информационно-поисковые системы: учеб.-метод. пособие. СПб.: СПбГУ, 2005б.

20. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учеб. пособие. М.: Академия, 2004.

21. Клышинский Э.С. Начальные этапы анализа текста // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. С. 106—140.

22. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. М.: Академия, 2006.

23. Марчук Ю.Н. Проблемы машинного перевода. М.: Наука, 1983.

24. Марчук Ю.Н. Автоматический перевод // Большой энциклопедический словарь. Языкознание. М.: Большая Российская энциклопедия, 1998. С. 15.

25. Марчук Ю.Н. Компьютерная лингвистика: учеб. пособие. М.: АСТ Восток—Запад, 2007.

26. Марчук Ю.Н. Модели перевода. М.: Академия, 2010.

27. Мельчук И.А. Автоматический синтез // Большая советская энциклопедия. М.: Советская энциклопедия. 1969—1978. http://dic.academic.ru/dic.nsf/bse/61319/Автоматический (дата обращения: 28.02.2012).

28. Мечковская Н.Б. История языка и история коммуникации: от клинописи до Интернета: курс лекций по общему языкознанию. М.: Флинта: Наука, 2009.

29. Мыркин В.Я. Введение в языкознание. Архангельск: Поморский университет, 2005.

30. Овчинникова И.Г., Угланова И.А. Компьютерное моделирование вербальной коммуникации: учеб.-метод. пособие. М.: Флинта: Наука, 2009.

31. Ожегов С.И. Словарь русского языка. 22-е изд., стер. М.: Русский язык, 1990.

32. Орехов Б.В., Слободян Е.А. Проблемы автоматической морфологии аглютинативных языков и парсер башкирского языка // Информационные тех-

нологии и письменное наследие: материалы международной научной конференции (Уфа, 28—31 октября 2010 г.) / отв. ред. В.А. Баранов. Уфа; Ижевск: Вагант, 2010. С. 167—171.

33. Основы теории коммуникации: учебник / под ред. М.А. Василика. М.: Гардарики, 2007.
34. Потапова Р.К. Новые информационные технологии и лингвистика: учеб. пособие. 2-е изд. М.: Едиториал УРСС, 2004.
35. Потапова Р.К. Речевое управление роботом: лингвистика и современные автоматизированные системы. М., 2005.
36. Прикладное языкознание: учебник / отв. ред. А.С. Герд. СПб.: СПбГУ, 1996.
37. Рождественский Ю.В. Лекции по общему языкознанию. М.: Высшая школа, 1990.
38. Розина И.Н. Педагогическая компьютерно-опосредованная коммуникация: теория и практика. М.: Логос, 2005.
39. Русский ассоциативный словарь / Ю.Н. Караулов, Ю.А. Сорокин, Е.Ф. Тарасов. Кн. 5: Прямой словарь: от стимула к реакции. М.: ИРЯ РАН, 1998.
40. Селегей В. Электронные словари и компьютерная лексикография // Ассоциация лексикографов Lingvo. www.lingvoda.ru/transforum/articles/selegey_a1.asp (дата обращения: 28.02.2012).
41. Семенов А.Л. Современные информационные технологии и перевод. М.: Академия, 2008.
42. Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). М., 2005. www.aot.ru/docs/RusCorporaHMM.htm (дата обращения: 28.02.2012).
43. Степанов А.Н. Информатика: учеб. пособие. СПб.: Питер, 2006.
44. Сысоев П.В., Евстигнеев М.Н. Методика обучения иностранному языку с использованием новых информационно-коммуникационных Интернет технологий: учеб.-метод. пособие. М.: Глосса-Пресс; Ростов н/Д: Феникс, 2009.
45. Титова С.В. Ресурсы и службы Интернета в преподавании иностранных языков. М.: Изд-во МГУ, 2003.
46. Толдова С.Ю., Бонч-Осмоловская А.А. Автоматический морфологический анализ // Фонд знаний «Ломоносов». М., 2011. www.lomonosov-fund.ru/enc/ru/encyclopedia:0127430 (дата обращения: 28.02.2012).
47. Толдова С.Ю., Архипов А.В., Логинова Е.А., Попова Д.П. Корпусная лингвистика // Фонд знаний «Ломоносов». М., 2011. www.lomonosov-fund.ru/enc/ru/encyclopedia:01210:article (дата обращения: 28.02.2012).

48. Филиппович Ю., Чернышева М. Историческая компьютерная лексикография — terra incognita в компьютерном мире // Компьютерра. 1999. № 45. http://offline.computerra. ru/1999/323/ 3379 (дата обращения: 13.05.2011).

49. Фролов А.В., Фролов Г.В. Синтез и распознавание речи. Современные решения: электронный учебник. http://frolov-lib.ru/books/hi/ch00.html (дата обращения: 28.02.2012).

50. Хан У., Мани И. Системы автоматического реферирования // Открытые системы. 2000. № 12. www.osp.ru/os/2000/12/178370 (дата обращения: 02.02.2012).

51. Хомский Н., Миллер Дж. Введение в формальный анализ естественных языков. М.: Едиториал УРСС, 2003.

52. Хроленко А.Т., Денисов А.В. Современные информационные технологии для гуманитария: практ. руководство для студ., аспирантов, препод.-филологов. М.: Флинта: Наука, 2007.

53. Bolshakov I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. México, 2004.

54. Huang, C., Simon, P., Hsieh, S., & Prevot, L. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Word break Identification // Proceedings of the Association for Computational Linguistics. Demo and Poster Sessions. Prague, 2007. P. 69—72. http://www.aclweb.org/anthology/P/P07/P07-2018.pdf (дата обращения: 28.02.2012).

55. Kinnersley B. The Language List. Collected Information On About 2500 Computer Languages, Past and Present. http://people.ku.edu/~nkinners/LangList/Extras/langlist.htm (дата обращения: 28.02.2012).

56. Levy M. CALL: context and conceptualization. Oxford: Oxford University Press, 1997.

57. Maher J. Eliza // Digital Antiquaria, Interactive Fiction. June 15, 2011. http://www.filfre.net/2011/06/eliza-part-1/ (дата обращения: 28.02.2012).

58. Matthews C. Intelligent Computer Assisted Language Learning as cognitive science: the choice of syntactic frameworks for language tutoring // Journal of Artificial Intelligence in Education. 1994. № 5/4. P. 533—556.

59. Rayner M., Carter D. M., Bretan I., Eklund R., Wirén M., Hansen S.L., Kirchmeier-Andersen S., Philp C., Sorensen F., Erdman Thomsen H. Recycling Lingware in a Multilingual MT System // Computation and Language. 1997. www.aclweb.org/anthology/W/W97/W97-0910.pdf (дата обращения: 28.02.2012).

60. Villiger C. Lernsoftware // Angewandte Linguistik: Ein Lehrbuch / Hrsg. von K. Knapp, G. Antos, M. Becker-Mrotzek u.a. Tübingen; Basel: Francke Verlag, 2004. S. 187—206.

61. Warschauer M. Computer Assisted Language Learning: an Introduction // Multimedia language teaching / ed. by S. Fotos. Tokyo: Logos International, 1996. P. 3—20.

Интернет-ресурсы

1. Диалог: Международная русскоязычная конференция по компьютерной лингвистике. http://dialog-21.ru
2. Лаборатория компьютерной лингвистики Института проблем передачи информации РАН http://proling.iitp.ru/ru/node/1
3. Корпусная лингвистика. Машинный перевод. Прикладная лингвистика // Фонд знаний «Ломоносов». http://www.lomonosov-fund.ru/enc/ru/encyclopedia:01206:article
4. Корпусная лингвистика: тематический сайт СПбГУ и ИЛИ РАН. СПб., 2008. http://corpora.iling.spb.ru
5. Информационные технологии в филологии // Викиверситет. http://ru.wikiversity.org/wiki/Информационные_технологии_в_филологии
6. Компьютерная лингвистика: научно-образовательный портал «Лингвистика в России: ресурсы для исследователей». http://uisrussia.msu.ru/linguist/_B_comput_ling.jsp
7. Прикладная лингвистика: портал «Единое окно доступа к образовательным ресурсам». http://window.edu.ru/window/ catalog?p_rubr=2.2.73.12.15
8. Программы лингвистического анализа и обработки текста. http://asknet.ru/Analytics/programms.htm
9. Речевые технологии http://speech-soft.ru/index.php
10. Association for Computational Linguistics. http://www.aclweb.org
11. Cogprints: free software for Linguistics. University of Southampton. http://cogprints.org/view/subjects/ling.html
12. Computational linguistics: MIT Press Journal. http://www.mitpressjournals.org/loi/coli
13. Computer-Assisted Language Instruction Consortium. Texas State University. http://calico.org/
14. GATES: free software. The University of Sheffield, 1995—2011. http://gate.ac.uk
15. Information and Communications Technology for Language Teachers (ICT4LT). Slough, Thames Valley University. http://www.ict4lt.org/en/en_home.htm

16. Institut für Computerlinguistik an der Universität Heidelberg. URL: http://www.cl.uni-heidelberg.de
17. Language Technology World http://www.lt-world.org/
18. LINGUIST List. URL: http://linguistlist.org
19. Stanford Engineering Everywhere (SEE): Artificial Intelligence. Stanford University, 1997—2009. http://see.stanford.edu/see/courses.aspx

ПРИЛОЖЕНИЯ

Приложение 1

Глоссарий

Автоматический анализ звучащей речи: преобразование звучащей речи в печатный текст, над которым можно производить дальнейшие операции.

Автоматический синтез звучащей речи: процесс преобразования печатного текста, существующего в цифровой форме, в звучащий текст на естественном человеческом языке.

Автоматический анализ текста: последовательное преобразование текста на естественном человеческом языке, введенного в компьютер, в его лексемно-морфологические, синтаксические и семантические представления, понятные компьютеру.

Автоматический синтез текста: процесс преобразования лексемно-морфологических, синтаксических и семантических представлений в текст на естественном языке.

Автоматическое рабочее место лингвиста: совокупность аппаратных, программных и лингвистических средств, необходимых для автоматической обработки лингвистических данных.

Алгоритм: формализованное описание последовательности действий, приводящей к решению поставленной задачи.

Аннотация: краткое изложение содержания документа, дающее общее представление о его теме, т.е. в отличие от реферата выполняющее лишь сигнальную функцию (есть публикация на определенную тему).

Веб 2.0 (социальная сеть): социальные сервисы и службы Всемирной паутины (блоги, веб-квесты, вики-проекты и т.п.), позволяющие широкому кругу людей быть не только получателями информации, но и ее создателями и соавторами.

Веб-ресурс: электронный документ, содержащий информацию различного рода (вербальную, графическую, табличную, звуковую, графическую, видеофайлы, анимацию и компьютерные программы), доступную через веб-страницы, размещенные во Всемирной паутине.

Дистанционное обучение: форма организации учебного процесса, основывающаяся на принципе самостоятельного получения знаний, предполагающая телекоммуникационный принцип доставки учебного материала и интерактивное взаимодействие обучающихся и преподавателей.

Естественный язык: исторически сложившаяся и используемая в определенной этнической группе или национальном государстве знаковая система.

Информатика: наука о накоплении, обработке и передаче информации с помощью электронных вычислительных машин.

Информационно-поисковая система: упорядоченная совокупность документов и информационных технологий, предназначенных для хранения и поиска информации в виде целых текстов или отдельных упоминаемых в них фактов.

Информационные революции: принципиальные изменения в способах фиксации и передачи информации, связанные с изобретением новых технических средств.

Информационно-поисковый язык: формальный язык, предназначенный для описания содержания документов, хранящихся в информационно-поисковой системе, и запроса пользователя.

Информационные технологии: компьютерные инструменты получения, хранения, передачи, распространения и преобразования информации, а также соответствующие законы и методы.

Информационные технологии в лингвистике: компьютерные инструменты получения, хранения, передачи, распространения и преобразования информации о языке и законах его функционирования , а также соответствующие законы и методы.

Информационный шум: множество документов, выдаваемых в процессе информационного поиска, формально соответствующих запросу (релевантных), но не являющихся релевантными по смыслу.

Информация: сведения об окружающем мире, передаваемые человеком, живыми организмами или техническими системами для регулирования своего поведения в окружающей среде.

Искусственные языки: знаковые системы, искусственно создаваемые в тех областях, где применение естественных языков менее эффективно или невозможно.

Искусственный интеллект: направление в информатике, связанное с созданием сложных человеко-машинных и робототехнических систем.

Кибернетика: наука об управлении, связи и переработке информации.

Ключевое слово: слово, относящееся к основному содержанию текста и повторяющееся в нем несколько раз.

Кодирование: процесс представления информации в виде последовательности условных обозначений; сопоставление объектов и отношений между ними с символами или словами какого-либо языка.

Компьютер: электронное устройство, служащее для создания, обработки, передачи и воспроизводства информации по написанным человеком алгоритмам (программам).

Компьютерная лексикография: раздел прикладной лингвистики, нацеленный на создание электронных (автоматических) словарей, лингвистических баз данных и разработку программ поддержки лексикографических работ.

Компьютерная лингвистика: область использования компьютерных инструментов — программ, технологий организации и обработки данных — для моделирования функционирования языка в тех или иных условиях, а также сферу применения компьютерных моделей языка в лингвистике и смежных с ней дисциплинах.

Компьютерная терминография: наука о составлении электронных терминологических словарей.

Компьютерное обучение языкам (CALL): область знаний и практических действий, нацеленных на использование компьютеров в обучении и изучении языков, имеющая свою методику, программные средства, цели и задачи.

Корпус лингвистический: совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска.

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

Лемма: словарная форма лексемы.

Лингвистика: наука о закономерностях происхождения, строения и функционирования естественного человеческого языка.

Лингвистические ресурсы (lingware): грамматические справочники, словари, энциклопедии, лингвистические базы данных и другие ресурсы, существующие в цифровой форме, доступные для компьютерной обработки на компьютере пользователя или размещенные в Интернете.

Локализация: перевод веб-страниц и компьютерных программ; в последнем случае переводу подвергаются сообщения об ошибках, тексты меню и служебной информации и т.д., предназначенные для человека и распределенные внутри компьютерных программ.

Машинная основа: последовательность букв от начала словоформы, общая для всех словоформ, входящих в формообразовательную парадигму данного слова, например (рус.) блок#, включ# или (англ.) buil#, earl#.

Машинное окончание: элемент, описывающий формоизменение конкретной лексемы; машинные окончания представляются в виде парадигм.

Машинный перевод: передача содержания текста на одном языке средствами другого языка с использованием компьютеров.

Модель: материальный или идеальный образ некоторой совокупности предметов или явлений, заменяющий реальные предметы и явления и включающий их наиболее существенные признаки. В лингвистике модели имитируют строение и функционирование языка, производство и восприятие речи и текста.

Основа: ядерная часть слова без словоизменительных морфем.

Парсер: специальная компьютерная программа для автоматического анализа слов, морфологического или синтаксического.

Поисковая система: специальная программа, позволяющая находить веб-ресурсы, текстовое содержание которых соответствует запросу пользователя.

Прикладная лингвистика: область языкознания, связанная с разработкой методов решения практических задач использования языка; отвечает на вопрос «Как лучше использовать язык?».

Программа: созданный человеком алгоритм для автоматического выполнения компьютером действий над информацией различного рода.

Программа автоматического распознавания текста (OCR-программа): компьютерная программа, позволяющая преобразовать текст с бумажного носителя в электронный текстовый файл, который в дальнейшем может обрабатываться человеком в любом текстовом редакторе.

Разметка (tagging, annotation): приписывание текстам и их компонентам специальных меток (тэгов).

Реферат: связный текст, который кратко выражает центральную тему, предмет, цель, методы и результаты исследования; обычно составляется к научно-техническим документам: научным монографиям, статьям, патентам на изобретение и др.

Символ: знак, обозначающий некоторый предмет или явление; в информатике — любой знак (буква, цифра, знак препинания, пробел и т.д.).

Система переводческой памяти (Translation Memory, TM): программа, сохраняющая переводы, сделанные ранее, и предлагающая человеку уже готовый перевод фразы или фрагмента текста, если он уже был однажды переведен.

Снятие омонимии (=разрешение многозначности): выбор правильной интерпретации словоформы, допускающей несколько вариантов толкований.

Терминологические базы (банки) данных (ТБД): массивы терминов по одной или разным областям знания, сохраняемые в электронном виде и снабженные системами автоматического поиска.

Тест Тьюринга: тест, смысл которого сводится к констатации факта создания искусственного интеллекта: проблема создания искусственного интеллекта решена, если большинство участников общения не смогут установить, с кем они беседуют — с человеком или машиной.

Электронный (автоматический) словарь: собрание слов и их комментариев в специальном машинном формате, предназначенное для использования человеком или являющееся составной частью более сложных компьютерных программ (например, систем машинного перевода).

Язык: знаковая система, используемая для общения в некотором социуме.

Приложение 2

Темы докладов по курсу

1. Обзор сетевых ресурсов по корпусной лингвистике
2. Характеристика ресурсов по компьютерной лингвистике (www.dialog-21.ru, www.computer.org)
3. Специальные возможности программы MS Word для лингвистов (проверка правописания, рецензирование, автореферирование, использование шаблонов и т.д.)
4. Правильное использование заимствованных терминов и обозначений (правописание, склонение, спряжение, ударение) компьютерной лингвистики
5. Особенности электронных переводческих словарей *Lingvo* и *Multitran* и их отличия от онлайн-переводчиков (*Google*, *Yandex* и т.п.)
6. Сравнение программ переводческой памяти (*TRADOS*, *Déjà vu* и т.п.)
7. Сравнение программ автоматического перевода (*ПРОМТ*, *Сократ* и т.п.)
8. Средства обеспечения и поддержки локализации (*Multilizer*, *Passolo* и т.п.)
9. Краудсорсинг или модель «Википедии» в переводе
10. Сравнение мультимедийных программ по обучению иностранным языкам (*English DeLuxe*, «*РЕПЕТИТОР English*» и т.п.)
11. Технология подкастинга в обучении языкам
12. ВебКвесты в обучении языкам
13. Возможности электронного письма в обучении языкам
14. Сетевые формы коммуникации (электронная почта, чаты, форумы) и их влияние на язык
15. Ресурсы Всемирной паутины для обучения языкам

16. Сравнительный анализ составления поисковых запросов в популярных русскоязычных поисковых системах (*Google, Yandex, Rambler, Mail.ru, Altavista, Yahoo, MSN, AOL*)

Формальные требования. Работа представляется устно на семинаре (5—10 минут) и сдается для проверки преподавателю в электронном виде (презентация PowerPoint или документ MS Word, см. требования по форматированию доклада в формате MS Word в лабораторной работе 3).

Обязательные элементы электронного варианта работы:

- титульный слайд (страница): ФИО выступающего, группа, дата, тема, название курса и ФИО преподавателя;
- основной текст (5—10 слайдов или 2—4 страницы шрифтом Century Schoolbook, 12, 1,5 интервал, выравнивание по ширине); страницы должны быть пронумерованы, начиная с первой, в правом верхнем углу, но номер на 1-й странице не ставится (поставить соответствующую галочку в пункте меню «Формат номера страницы»);
- выводы (несколько ключевых предложений);
- список использованной научной литературы и/или сетевых ресурсов (от 2 до 10 наименований).

Технология подготовки доклада. Студент готовит доклад, чтобы продемонстрировать умение самостоятельно подбирать литературу по заданной теме, обрабатывать ее, ясно излагать полученное содержание устно и письменно. Этапы подготовки доклада:

1) Студент выбирает тему и согласовывает с преподавателем дату будущего выступления.
2) Студент самостоятельно или после консультации с преподавателем подбирает литературу по теме и необходимые Интернет-ресурсы, изучает их.
3) Студент внимательно изучает собранную литературу и обрабатывает ее: составляет конспект, выделяет ключевые идеи, пересказывает основное содержание прочитанного, при необходимости выбирает наиболее важные фрагменты для оформления цитат, сопоставляет разные мнения, оценивает и обобщает прочитанное. На этом этапе студентом создается собственный текст доклада, оформленный в виде документа MS Word и/или презентации PowerPoint.
4) Устное выступление: представление доклада и ответы на вопросы аудитории.
5) Окончательное оформление электронного варианта доклада с учетом заданных вопросов и сдача работы преподавателю.

Приложение 3

Тест для проверки знаний по курсу

Время на выполнение теста: 45 минут
В каждом задании — 1 правильный ответ,
за каждый правильный ответ дается 1 балл

1. Какое из высказываний является определением прикладной лингвистики?
   a) область языкознания, направленная на объективное установление состояния отдельного языка, его истории и закономерностей;
   b) область языкознания, связанная с использованием компьютерных инструментов — программ, технологий организации и обработки данных — для моделирования функционирования языка в тех или иных условиях;
   c) область языкознания, связанная с разработкой методов решения практических задач использования языка;
   d) область языкознания, связанная с применением компьютерных моделей языка в лингвистике и в смежных с ней дисциплинах.
2. К направлениям компьютерной лингвистики не относится
   a) компьютерная лексикография;
   b) компьютерно-опосредованная коммуникация;
   c) системы обработки естественного языка;
   d) машинный перевод.
3. Информатика — это
   a) наука об управлении, связи и переработке информации;
   b) наука о накоплении, обработке и передаче информации с помощью ЭВМ;
   c) наука о накоплении, обработке и передаче информации о строении языка с помощью ЭВМ;
   d) наука об использовании компьютерных инструментов для моделирования функционирования языка в тех или иных условиях.
4. Разное количество информации в одном и том же сообщении для разных людей зависит не от...
   a) накопленных ими знаний;
   b) уровня понимания сообщения;

c) их интереса к сообщению;

d) их уровня владения компьютерной техникой.

5. Следствие третьей информационной революции состоит в том, что...

a) информация становится общедоступной;

b) информацию можно автоматически обрабатывать и передавать с большой скоростью;

c) информацию можно легко найти с помощью инструментов поиска и совместно производить;

d) информация может накапливаться.

6. Для современного человека преобладающей является...

a) звуковая информация;

b) визуальная (символьная) информация;

c) вкусовая и тактильная информация;

d) визуальная (образная) информация.

7. Адекватность информации — это ...

a) степень соответствия информации объективной реальности окружающего мира;

b) степень соответствия информации, полученной потребителем, тому, что автор вложил в ее содержание;

c) достаточность информации для принятия решения;

d) степень соответствия информации текущему моменту времени.

8. Машинный синтаксис — это ...

a) правила строения имен;

b) правила построения слов в более сложные структуры;

c) соотношение слова и его значения;

d) правила перевода письменного символа в устный.

9. Естественный язык — это ...

a) знаковая система, используемая человеком с момента рождения;

b) знаковая система, используемая человеком в непринужденной обстановке;

c) знаковая система, созданная для естественных наук;

d) знаковая система, стихийно возникшая и закрепившаяся в обществе.

10. Волапюк — это ...
    a) специализированный язык науки;
    b) родной язык одного из малочисленных племен;
    c) неспециализированный искусственный язык;
    d) система символического кодирования.
11. Какие из следующих приложений не являются текстовыми редакторами?
    a) MS Excel;
    b) Corel WordPerfect;
    c) MS Works;
    d) Adobe InCopy.
12. Microsoft Word не включает...
    a) функции настольных издательских систем;
    b) функцию удалённого доступа;
    c) функцию редактирования графических объектов;
    d) шаблоны типовых таблиц.
13. К устройствам ввода данных не относится
    a) сканер;
    b) принтер;
    c) клавиатура;
    d) цифровой фотоаппарат.
14. OCR — это ...
    a) система автоматического распознавания символов;
    b) система переводческой памяти;
    c) система машинного перевода;
    d) функция текстового процессора.
15. Реферат — это ...
    a) связный текст, который кратко выражает тему, предмет, цель, методы и результаты исследования;
    b) процесс составления содержания документа (книги, статьи, патента на изобретение и др.);
    c) краткое изложение содержания документа, дающее общее представление о его теме;

d) краткий текст, выполняющий сигнальную функцию (информирует о том, что есть публикация на определенную тему).

16. Слово, относящееся к основному содержанию текста и повторяющееся в нем несколько раз, в автоматическом реферировании называется ...

a) лейтмотивом;

b) термином;

c) символом;

d) ключевым словом.

17. Метод автоматического аннотирования, при котором важные слова выделяются в заголовке, подзаголовке, начале и конце текста, называется ...

a) статистическим;

b) логико-семантическим;

c) позиционным;

d) функциональным.

18. Совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска, называется ...

a) базой данных;

b) словарем;

c) информационным массивом;

d) корпусом.

19. Разметка бывает ...

a) морфологической; синтаксической; семантической и просодической;

b) полнотекстовой и фрагментной;

c) синхронической и диахронической;

d) звуковой, письменной, смешанной.

20. УНК — это ...

a) корпус естественного языка, представительный по отношению ко всему языку;

b) универсальный национальный код;

c) собрание текстов, которое существует в Интернете;

d) собрание текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска.

21. Требования к корпусам

    a) полнота, адекватность, актуальность, компьютерная поддержка;

    b) устойчивость, тиражируемость, адаптируемость, оптимальность временных параметров, комфорт пользователя;

    c) репрезентативность, полнота, экономичность, структуризация, компьютерная поддержка;

    d) полнота, экономичность, достоверность, структуризация, компьютерная поддержка.

22. Корпусный менеджер ...

    a) обеспечивает сортировку результатов поиска, статистические подсчеты, составление списков слов на основе корпуса;

    b) это специальная программа поиска по корпусу;

    c) это человек, составляющий корпуса и управляющий ими;

    d) это специальная программа подготовки текстов к их включению в корпус.

23. ПОД — это ...

    a) вид информационно-поисковой системы;

    b) специальная программа поиска по корпусу;

    c) поисковый образ документа;

    d) поисковая оценка данных.

24. Одна из основных проблем компьютерного анализа речи состоит в том, что ...

    a) невозможно создать искусственный интеллект;

    b) компьютер не умеет работать со смыслом;

    c) у компьютера нет дополнительных источников информации (ситуация, контекст, прошлый опыт в данной области и т.п.);

    d) разработчики не желают делиться своими профессиональными секретами.

25. Электронный словарь — это ...

    a) введенный в компьютер бумажный словарь, снабженный средствами поиска и отображения информации;

    b) организованное собрание слов с комментариями, в которых описываются особенности структуры и/или функционирования этих слов;

c) организованное собрание слов с описанием их значения, особенностей употребления, структурных свойств, сочетаемости, соотношения с лексическими системами других языков и т.д.;

d) словарь в специальном машинном формате, предназначенный для применения на ЭВМ пользователем или компьютерной программой.

26. К зонам словарной статьи не относится

a) лексический вход (вокабула, лемма);

b) зона грамматической информации;

c) зона стилистических помет;

d) словник.

27. Что включает в себя понятие АСПОТ?

a) словарь в специальном машинном формате, предназначенный для применения на ЭВМ пользователем;

b) компьютерные версии хорошо известных словарей (Вебстер, Коллинз, Ожегов...);

c) словарь в специальном машинном формате, предназначенный для применения на ЭВМ компьютерной программой;

d) словари, предназначенные для обычного пользователя.

28. Что не относится к понятию термина?

a) слово (словосочетание) метаязыка науки, а также областей конкретной практической деятельности человека;

b) понятие задается через свойства, реализуемые в системе;

c) использование основывается не на интуиции, а на четких определениях;

d) сопоставляется, как правило, несколько значений.

29. Что не относится к процессу и понятию машинного перевода?

a) междисциплинарность;

b) использование машинных средств;

c) принципиальное сходство этапов понимания и синтеза текста;

d) учет языковых и экстралингвистических знаний.

30. Типовая парадигма лексемы в автоматическом морфологическом анализе — это ...

a) последовательность букв от начала словоформы, общая для всех словоформ;

b) элементы, описывающие формоизменение конкретной лексемы,

c) совокупность наборов машинных окончаний;

d) совпадение основ разных слов.

31. Требования к системам МП включают ...

a) устойчивость, тиражируемость, адаптируемость, оптимальность временных параметров, комфорт пользователя;

b) полнота, адекватность, актуальность, достоверность;

c) репрезентативность, полнота, экономичность, адекватность, компьютерная поддержка;

d) репрезентативность, полнота, экономичность, структуризация, компьютерная поддержка.

32. Аббревиатура CALL относится к ...

a) науке об использовании компьютерных инструментов для моделирования функционирования языка в тех или иных условиях;

b) обучению иностранному языку;

c) обучению языку с помощью компьютера;

d) использованию компьютеров в обучении.

33. Сущность когнитивно-интеллектуального подхода в компьютерном обучении состоит в том, что ...

a) программы ориентированы на обучающегося, дают свободу выбора уровня и типа действий;

b) программы построены по формуле стимул — реакция;

c) обучающемуся отводится роль объекта обучения;

d) в нем используются программы-тренажеры обучению языку с помощью компьютера.

34. К обучающим программным средствам не относятся ...

a) тестирующие программы;

b) энциклопедии;

c) программы-ассемблеры;

d) учебные игры.

35. Компьютерный учебник — это ...

a) программа, предлагающая пользователю вопрос и несколько вариантов ответов на него;

b) программа формирования автоматического навыка выполнения определенных коммуникативных действий путем многочисленных повторов;

c) программы, предназначенные для представления учебного материала;

d) программно-методический комплекс, позволяющий самостоятельно освоить учебный курс или его большой раздел.

36. Что не относится к компьютерным обучающим программам?

a) заменяют преподавателя;

b) организация и выполнение рутинной работы;

c) повышение активности обучаемого;

d) создание возможностей для самообразования.

Критерии оценки

32—36 баллов — отлично,

21—31 балл — хорошо,

15—20 баллов — удовлетворительно,

0—14 баллов — неудовлетворительно.

Приложение 4

Ключи к тесту

1 c, 2 b, 3 b, 4 d, 5 b, 6 b, 7 b, 8 a, 9 d, 10 c, 11 a, 12 b, 13 b, 14 a, 15 a, 16 d, 17 c, 18 d, 19 a, 20 a, 21 c, 22 a, 23 c, 24 b, 25 d, 26 d, 27 c, 28 d, 29 c, 30 c, 31 a, 32 c, 33 a, 34 c, 35 d, 36 a.

Учебное издание

Щипицина Лариса Юрьевна

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ЛИНГВИСТИКЕ

Учебное пособие

Подписано в печать 29.08.2012. Формат 60 × 88/16. Печать офсетная.
Усл. печ. л. 7,84. Уч.-изд. л. 5,21. Тираж 500 экз. Изд. № 2574.

ООО «ФЛИНТА», 117342, г. Москва, ул. Бутлерова, д.17-Б, комн. 324.
Тел./факс: (495)334-82-65, тел. (495)336-03-11.
E-mail: flinta@mail.ru; WebSite: www.flinta.ru.

Издательство «Наука», 117997, ГСП-7, Москва В-485,
ул. Профсоюзная, д. 90.