МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Е. П. Соснина

ВВЕДЕНИЕ В ПРИКЛАДНУЮ ЛИНГВИСТИКУ

**Учебное пособие по курсу
«Основы теоретической и прикладной лингвистики»
для студентов направления «Лингвистика»**

Ульяновск
УлГТУ
2012

УДК 81'33 (075)
ББК 81.1я73
С 66

Рецензенты:

доктор филологических наук, профессор,
зав.кафедрой общего и германского языкознания
Ульяновского государственного университета
А. И. Фефилов;

доктор филологических наук, профессор,
зав.кафедрой Прикладной лингвистики
Пензенской государственной технологической
академии *Д. Н. Жаткин.*

*Утверждено редакционно-издательским советом
университета в качестве учебного пособия*

Соснина, Е. П.
С66 Введение в прикладную лингвистику : учебное пособие / Е. П. Соснина. – 2-е изд., испр. и доп. – Ульяновск : УлГТУ, 2012. – 110 с.

ISBN 978-5-9795-1018-7

Составлено с целью помочь студентам направления «Лингвистика» в ознакомлении с основными задачами прикладной лингвистики и затрагивает такие важные темы, как «Теоретическая и прикладная лингвистика», «Моделирование в лингвистике и искусственные языки», «Машинный перевод», «Лингвистика в задачах информационного поиска», «Компьютерная лингвистика» и др. Каждая тема включает конспект лекции, программу семинарских занятий, список литературы, домашние задания. В приложениях даны рабочая программа по курсу, вопросы для подготовки к экзаменам, примеры лингвистических задач, основная терминология.

Работа подготовлена на кафедре «Прикладная лингвистика» УлГТУ.

УДК 81'33 (075)
ББК 81.1я73

ISBN 978-5-9795-1018-7

ВВЕДЕНИЕ

Данное издание является второй существенно дополненной редакцией вышедшего в 2000 году пособия «Введение в прикладную лингвистику» с авторскими конспектами лекций доцента УлГТУ Сосниной Е. П.

Настоящее учебное пособие состоит из конспектов восьми лекций, списка литературы, практических приложений. После каждого лекционного раздела приводится тематика семинаров, список домашних заданий. Последнее приложение дает двуязычный список основных терминов, встречаемых в лекциях и обсуждаемых на семинарских занятиях.

Целью курса «Основы теоретической и прикладной лингвистики», для методической поддержки которой и разрабатывалось данное учебное пособие, является получение базовых знаний в области прикладного языкознания и ознакомление студентов-лингвистов начального курса обучения с основными направлениями прикладной лингвистики (машинный перевод, информационный поиск, компьютерная лингвистика, квантитативная лингвистика и т.д.) и междисциплинарной связью лингвистики с другими науками. Программа акцентирует внимание студентов на современных приложениях лингвистики и ее связи с новыми информационными технологиями. Кроме того, изучение дисциплины служит целям формирования мировоззрения, развития интеллекта, эрудиции, формирования профессиональных компетенций по специальности.

Основные ***задачи курса***: ознакомить студентов с современным понятийным аппаратом, терминологией, современными подходами, методами и инструментарием, технологиями в области прикладной лингвистики в России и за рубежом, вариантами искусственных языков, методами их моделирования и представления, лингвистическим обеспечением различных компьютерных систем (автоматической обработки языка и речи, информационно-поисковых, экспертных систем, лексикографических, систем машинного перевода и др.); объяснить взаимосвязь и взаимодействие теоретической и прикладной

лингвистики, а также других смежных наук и направлений; научить решать лингвистические задачи, связанные с моделированием элементов искусственных и естественных языков.

Краткая характеристика дисциплины, ее место в учебном процессе

Дисциплина закладывает базовые знания специализации лингвиста в области новых компьютерных и информационных технологий, которые будут конкретизированы и дополнены содержанием последующих общепрофессиональных и специальных дисциплин.

Требования к уровню освоения дисциплины

***Основные требования*:** студенты должны четко представлять структурные уровни языка, возможности моделирования и создания искусственных языков. Все семинарские занятия рассчитаны на решение лингвистических задач по пройденной теме, а также конспектирование обязательной литературы.

В результате изучения дисциплины:

- студент должен ***знать*** современные направления развития прикладного языкознания, о лингвистическом обеспечении систем обработки языка и речи; обладать базовыми знаниями в области структурного описания языков (по уровням); иметь четкое представление о структуре лингвистики как науки, о связи ее с другими научными дисциплинами, в частности, информатикой, о пограничных дисциплинах (логика, семиотика, психолингвистика, социолингвистика, лингводидактика, теория и практика перевода и др.);

- студент должен ***уметь*** оперировать базовым понятийным аппаратом теоретической и прикладной лингвистики, решать практические задачи моделирования языковых аспектов на примере некоторых искусственных языков (языков формальных грамматик, форм Бэкус-Наура, поисковых, языков представления знаний и семантики);

- студент должен приобрести ***навыки*** самостоятельной, поисковой и исследовательской работы над учебным и научным материалом по пройденным

темам курса, а также практической работы с лингвистическими ресурсами (системами машинного перевода, словарями, тезаурусами, поисковыми системами, с корпусами, электронными учебными и тестовыми программами).

Методические рекомендации студентам

В силу довольно большого объема нового материала и различных взглядов на организацию учебных курсов по прикладной лингвистике в процессе изучения данной дисциплины следует опираться на конспекты лекций, изложенные в данном пособии. Другая основная и дополнительная литература доступна в библиотеке вуза, а также в открытом Интернет-доступе, на что в пособии даются соответствующие ссылки.

Для самостоятельной работы рекомендуется индивидуально прорешать лингвистические задачи, представленные как типовые на семинарских занятиях, и выполнить домашние задания.

Для более глубокого освоения и понимания тем дисциплины рекомендуется основное внимание уделять дополнительной и новейшей литературе на тематических порталах и сайтах, посвященным проблемам современной прикладной лингвистики, например, www.dialog_21.ru, и на открытом ресурсе http://books.google.com/. Литература, на которую ссылается автор, также представлена после каждого раздела-лекции.

На начальном этапе обучения введение в теорию языкознания дается студентам как отдельный раздел вводного модуля в параллельно проводимой дисциплине. Методической поддержкой теоретического раздела модуля является учебник профессора Шарафутдиновой Н. С. *Теория и история лингвистической науки* : учебник / *Н. С. Шарафутдинова*. – 3-е изд., испр. и доп. – Ульяновск : УлГТУ, 2012.

Учебно-методической поддержкой вводного раздела по прикладной лингвистике является данное учебное пособие.

Формы и методика текущего, промежуточного и итогового контроля

Текущий контроль проводится на каждом занятии (на лекции, в ходе семинарского занятия) и выполняет целый ряд функций: проверочную, оценочную, корректирующую, мотивирующую и дисциплинирующую. Контролируется выполнение конспектов и своевременное решение домашних задач.

Промежуточный контроль является тематическим. Проводится в форме промежуточного теста в период вузовской промежуточной аттестации (середина семестра) по завершении первых четырех тематических разделов программы и освоении определенного объема теоретического материала.

Итоговый контроль проводится в конце теоретического курса обучения по дисциплине. Форма – **итоговый тест** по дисциплине по всем тематическим разделам курса (примеры тестовых вопросов даны в конце пособия) и **устный экзамен** с обязательным решением одной лингвистической задачи и устным ответом на теоретические вопросы экзаменатора.

Некоторые комментарии автора

Хотелось бы выразить свою признательность своим читателям, коллегам из УлГТУ, МГУ, МГЛУ, РГГУ СПбГУ и др., коллективу кафедры «Прикладная лингвистика», а также студентам-лингвистам, которые внимательно слушали мои лекции по прикладной лингвистике, часто задавали неожиданные глобальные вопросы из разряда «А зачем ЭТО ВСЁ нам нужно?», временами ставя меня в тупик. Но, надеюсь, что со временем эта мозаика из задач, направлений и подходов прикладной лингвистики сложилась у них в красивую сложную картину, а каждый из моих бывших студентов нашёл своё достойное место в жизни.

Автор

КОНСПЕКТЫ ЛЕКЦИЙ ПО КУРСУ «ОСНОВЫ ТЕОРЕТИЧЕСКОЙ И ПРИКЛАДНОЙ ЛИНГВИСТИКИ»

Лекция 1

План:

I. Лингвистика как наука. Сложная система лингвистических дисциплин и направлений.

II. Теоретическая и прикладная (практическая) лингвистика. Задачи и направления прикладной лингвистики.

III. Связь лингвистики с другими науками – естественными и гуманитарными.

I. Большой энциклопедический словарь «Языкознание» [1] дает следующее определение лингвистической науке: *«**Лингвистика** (языкознание, языковедение) – это наука о естественном человеческом языке **вообще** и обо всех языках мира **как индивидуальных его представителях**».* Лингвистика в широком смысле занимается познанием языка и передачей результатов этого познания другим людям для их практических целей.

Интересен факт, что сам термин *«лингвистика» (linguistics)* вошел в научный обиход в 1847 году, хотя термин *«linguist*» (в значении *«a student of language*») возник на 200 лет раньше [2].

Языковедение возникло в связи с практическими потребностями людей и тесно связано с появлением письменности в период, по разным оценкам, от 2 до 5 тысяч лет до нашей эры на Древнем Востоке. Одними из первых лингвистических продуктов ученые считают шумерские глоссы (25 в. до н. э.), а первым серьезным научным трудом – формальные грамматики санскрита древнеиндийского ученого Панини (V в. до н. э.). Само же языковедение стало оформляться как отдельная наука лишь в XVII–XVIII веках, что было обусловлено ее огромной сложностью и недостатком знаний о языке – объекте этой науки. В XIX–XX вв. отмечается подъем в области языкознания, появляются известные лингвистические школы (американская, пражская,

копенгагенская, московская и др.) и направления. Огромный вклад в лингвистическую теорию вносит Фердинанд Де Соссюр, который разрабатывает основы структурализма, или структурную лингвистику, развивает теорию языка и речи [Шарафутдинова, 2012, с. 249-255]. Практическое направление, называемое сегодня «прикладная лингвистика», активно стало развиваться в середине XX века в связи с появлением компьютерных технологий и усовершенствованием технических средств.

История лингвистики как науки, различные лингвистические школы, этапы развития лингвистических учений рассматриваются подробно в рамках курса «История лингвистических учений» [Шарафутдинова, 2012, часть 2].

Языкознание изучает различные стороны языка, рассматривая их в рамках своих отдельных научных направлений. Например:

Фонетика и ***фонология*** – научные направления языкознания, изучающие звуковую систему языка и особенности ее функционирования.

Лексикология изучает словарный состав языка и закономерности его развития.

Грамматика, как наука о способах и средствах построения и изменения слова, о способах и средствах построения предложения, подразделяется на две области – морфологию и синтаксис. ***Морфология*** – учение о способах и средствах построения и изменения слов. ***Синтаксис*** – учение о способах и средствах построения предложений.

Семантика – наука о смысле (значениях) слов и предложений, частей речи и членов предложений.

Прагматика – учение об условиях и целях коммуникации, влияющих на понимание, так называемое «изучение языка в контексте», изучение отношений между средствами языка и теми, кто этими средствами пользуется.

Дискурс изучает и анализирует характерные для данного вида текстов (или коммуникантов) особенности лексики, синтаксиса, семантики либо прагматики языковых единиц, проявляющиеся в актуальных коммуникативных актах, речи и письменных текстах (например, спортивный дискурс).

Такой подход традиционен в изучении языка и нам частично он знаком из школьного курса преподавания русского и иностранных языков. Углубленно такой подход к структуре языка и теории языкознания дается и в вузе в рамках образовательных программ по лингвистике. Но мы также будем рассматривать лингвистику как сложную и разветвленную систему дисциплин как теоретического, так и прикладного характера. Приведем лишь некоторые из таких дисциплин для демонстрации разнообразия языковедческих направлений.

***Примеры лингвистических дисциплин*:**

1. *Общее языкознание* занимается проблемами, касающимися всех языков, исследует сущность и природу языка вообще, проблему его происхождения и общие законы его функционирования, структуру, классификацию языков. Общее языкознание разрабатывает методы исследования языков, формулирует языковые универсалии, т.е. положения, действительные для всех или большинства языков мира.

2. *Частное языкознание* изучает один или группу родственных языков (например, германистика, романистика, тюркология, балканистика и т.д.). В зависимости от целей и задач исследования частное языкознание может быть *синхроническим*, описывающим фактор языка в какой-либо момент истории, в одной временной плоскости, чаще всего современное его состояние, и *диахроническим*, прослеживающим историческое развитие языка.

3. *Сравнительно-историческое языкознание* (компаративистика) занимается установлением степени родства между языками (построением генеалогической классификации языков), реконструкцией первых языков, или праязыков, исследованием различных процессов в истории языков, их групп и семей, этимологией слов.

4. *Типология* (универсализм) занимается выяснением наиболее общих закономерностей различных языков, не связанных между собой общим происхождением или взаимовлиянием. В случае, если некоторое явление выявляется в представительной группе языков, оно может считаться типологической закономерностью, применимой к языку как таковому.

Типологический анализ возможен на уровне звука (фонетическая типология), на уровне слова (морфологическая типология), предложения (синтаксическая типология) и далее (типология текста или дискурса).

5. *Диалектология* (ее вариант – *лингвистическая география*) – изучает местные территориальные разновидности одного языка.

6. *Когнитивная лингвистика* – междисциплинарное направление языкознания, тесно связанное с семантикой, которое исследует проблемы соотношения языка, сознания и мышления, роль языка в познании, понимании и отражении окружающей действительности.

7. *Паралингвистика* – изучает неязыковые (невербальные) средства в речи, передающие совместно с вербальными смысловую информацию в составе речевого сообщения, а также совокупность таких средств. Изучение невербального общения является важным компонентом межкультурной коммуникации. Паралингвистику также считают разделом невербальной семиотики.

8. *Психолингвистика* – связана с изучением речи, речеобразования (чаще всего конкретного индивида) для решения тех проблем, которые не могут быть решены ни в психологии, ни в лингвистике. Психолингвистика – самостоятельное научное направление, возникшее на стыке лингвистики и психологии. Объединение этих двух дисциплин в одну пограничную дисциплину позволяет использовать понятийный аппарат лингвистики для описания языковой формы речевых высказываний и понятийный аппарат психологии для описания и объяснения психических процессов производства и восприятия речи. В рамках психолингвистики изучается, например, детская речь, речевые отклонения при психических заболеваниях.

9. *Социолингвистика* – изучает роль языка в обществе, воздействие общества на язык, язык в связи с социальными условиями его существования, комплексом внешних обстоятельств, в которых реально функционирует и развивается язык: общество людей, использующих данный язык, социальная структура этого общества, различия между носителями языка в возрасте,

социальном статусе, уровне культуры и образования, месте проживания, а также различия в их речевом поведении в зависимости от ситуации общения. В рамках социолингвистики активно развивается, например, *политическая лингвистика*.

10. *Прикладная лингвистика* – к этой области относятся практические направления языкознания, часто междисциплинарного характера, например, *компьютерная лингвистика* (машинный перевод, лингвистические основы информатики, информационный поиск и др.), *лингводидактика* (обучение языку), *дешифровка* (исследование текстов на незнакомом коде, или языке, для получения информации), *математическая лингвистика* (например, квантитативная лингвистика, проводящая статистический анализ языка и текстов, занимающаяся построением и использованием структурно-вероятностных моделей языков).

II. Теоретическая и прикладная лингвистика

Лингвистику условно принято подразделять на *теоретическую*, иногда ее еще называют *«научная лингвистика»,* или *«теория языкознания»* (в рамках этого направления рассматриваются различные научные концепции, лингвистические теории, лингвистические школы, язык с точки зрения его структуры и систем), и *прикладную* (практическую) лингвистику.

Одним из важнейших направлений развития лингвистики является ***прикладная лингвистика***, одной из главных задач которой является решение и оптимизация традиционных лингвистических задач, как, например, перевод или обучение языкам. Обеспечению практических потребностей людей служат такие разделы прикладной лингвистики, как написание практических грамматик, практическая фонетика, практическая лексикография (создание учебных или отраслевых словарей).

Каждое из направлений практической лингвистики обычно имеет свое отражение в сфере теоретической лингвистики. Так, например, проблемы перевода изучаются в рамках такой дисциплины как переводоведение

(Translation Studies), которая формирует свою теорию перевода, различные переводческие концепции.

В российской традиции изначально было принято широкое понимание термина **«прикладная лингвистика»**, на что указывает профессор А. Н. Баранов в своем вводном курсе и пособии «Введение в прикладную лингвистику» [Баранов, 2003]. И в такой широкой интерпретации *прикладная лингвистика* является комплексной научной дисциплиной, применяющей лингвистические знания в различных ситуациях для решения разного рода практических задач (таких, например, как машинный перевод, техническая коммуникация, распознавание и синтез речи, информационный поиск и др.).

Следует также отметить, что за рубежом термин **«Applied Linguistics»** иногда трактуется узко, а под ним часто понимается только сфера преподавания иностранных языков, т. е. сфера подготовки преподавателей иностранных языков, изучение методики лингводидактики. Вот что пишет о происхождении указанного значения термина Американское лингвистическое общество: *«Applied linguistics ...was first officially recognized as an independent course at the University of Michigan in 1946. In those early days, the term was used both in the United States and in Great Britain to refer to applying a so-called 'scientific approach' to teaching foreign languages, including English for nonnative speakers. Early work to improve the quality of foreign language teaching ...helped to bring definition to the field as did the 1948 publication of a new journal, Language Learning: A Quarterly Journal of Applied Linguistics».*

В настоящее время Американское лингвистическое общество также определяет прикладную лингвистику как более широкую область знаний: «*Applied linguistics refers to a broad range of activities which involve solving some language-related problem or addressing some language-related concern».* Такое понимание термина имеет исторические корни: *«During the late 1950s and the early 1960s, the use of the term was gradually broadened to include what was then referred to as 'automatic translation'»* (см. http://lsadc.org).

По мнению многих ученых настоящий период развития языкознания характеризуется серьезным влиянием на лингвистическую теорию прикладной лингвистики. По словам известного ученого В. А. Звегинцева, прикладная лингвистика является *«полигоном, где проходят испытания как частные гипотезы, так и глобальные теоретические построения…»* [Звегинцев, 2008]. Также, когда говорят о проверке лингвистических теорий на практике, то имеют в виду не только, например, **создание** практических грамматик и словарей, но и их эффективное **приложение** (использование) в обучении и преподавании.

Современная прикладная лингвистика активно развивается и решает многие проблемы обработки устной и письменной речи, извлечения и использования информации из текстов, повышает эффективность коммуникации, в том числе в сфере информационных и компьютерных технологий.

Прикладная лингвистика представляет собой ту сферу, где реально проводятся лингвистические эксперименты, имеющие целью верификацию положений теории языкознания и проверку эффективности лингвистических продуктов, создаваемых разработчиками. Как говорит профессор СПбГУ А. С. Герд: *«К счастью, прикладная лингвистика – наука экспериментальная*» [Герд, 2005, стр. 5].

Задачи и направления прикладной лингвистики

Определим круг некоторых *основных задач* прикладной лингвистики:

1. Перевод с/на иностранный язык;
2. Обучение иностранному языку (методики обучения), лингводидактика;
3. Поддержка коммуникации с помощью технических средств;
4. Лингвистические основы информатики (создание искусственных языков, например, языков программирования, web-разработки);
5. Лингвистические основы информационного поиска;
6. Аннотирование, реферирование, классификация текстов;

7. Составление словарей (практическая лексикография);
8. Терминология и терминография (упорядочение, стандартизация и унификация научно-технической терминологии);
9. Распознавание символов текста;
10. Распознавание и синтез устной речи.

В последние 50 лет внедрения новых информационных технологий во все сферы человеческого общения прикладная лингвистика развивается по направлению автоматизации основных задач, оптимизации коммуникации, т. е. развивается такое направление прикладной лингвистики, которое назвали **компьютерной лингвистикой**. Сюда входят, например:

- Машинный перевод (Machine Translation);
- Компьютерная лексикография (подготовка электронных словарей);
- Компьютерная лингводидактика (CALL/T – Computer Assisted Language Learning and Teaching);
- Автоматическая обработка естественных языков (NLP – Natural Language Processing).

В рамках прикладной лингвистики для решения вышеуказанных задач необходимо было выработать эффективные **методы исследования языка**. Среди таких классических методов можно отметить, например, вероятностно-статистические методы, логико-математические методы, методы моделирования, корпусного анализа, контент-анализа и многие другие, которые активно и плодотворно использовались в других науках и нашли свое эффективное приложение в лингвистике. Некоторые из методов мы рассмотрим ниже в других разделах нашего курса. Особое внимание мы уделим методу моделирования.

II. Лингвистика – сложная многоплановая область знаний и приложений, которая соприкасается как со многими гуманитарными науками, так и с естественнонаучными. В последние годы появились смежные лингвистические

направления на стыке тех или иных наук, которые активно развиваются за счет новых теоретических и практических приложений. Укажем лишь некоторые:

- Социология, Политология - > *Социолингвистика, Политическая лингвистика;*
- Филология - > *Лингвистика текста*;
- Психология - > *Психолингвистика*;
- Логика - > *Когнитивная лингвистика, Информационный поиск, Моделирование семантики и знаний*;
- Математика - > *Математическая лингвистика, Лингвостатистика*;
- Биология - > *Фонетика;*
- Физика - > *Распознавание и синтез речи;*
- История, археология этнография - > *Этнолингвистика, Дешифровка, Сравнительно-историческое языкознание;*
- Юридическая наука - > *Юридическая лингвистика* (рассматривает язык права, проводит лингвистические экспертизы текстов и высказываний);
- Информатика - > *Компьютерная лингвистика* (например, ее раздел – лингвистические основы информатики, связанный с разработкой и обработкой искусственных языков программирования);
- Семиотика - > *Знаковая теория языка, WEB-дизайн*.

В следующей лекции мы более подробно остановимся на связи лингвистики и информатики, а также такой интересной науки, как семиотика, так как, по нашему мнению, эти науки наиболее актуально вписываются в круг интересов прикладной лингвистики.

Семинар

План семинара

I. *Речь и язык. Предмет и объект языкознания. Понятие о языке как средстве коммуникации.*
II. *Система и структура. Языковые системы. Структурные уровни языка.*

Домашнее задание

Сделать конспект главы 1 «Объект и методы прикладной лингвистки» по учебнику Баранов А. Н. Введение в прикладную лингвистику: Учебное пособие. – 2-е изд., испр. – М.: Едиториал УРСС, 2009. – 360 с.

Список литературы и электронных ресурсов

1. Большой энциклопедический словарь. Языкознание / под ред. В. Н. Ярцевой – М. : Большая Российская энциклопедия, 1998. – 685 с. (см. соответствующие разделы и определения под терминами – социолингвистика, психолингвистика и т. д.)
2. Этимологический словарь – Online Etymological Ductionary – http://www.etymonline.com/index.php?term=linguist (дата обращения: 17.07.2012)
3. Баранов, А. Н. Введение в прикладную лингвистику : учебное пособие. – 2-е изд., испр. – М. : Едиториал УРСС, 2009. – 360 с. (см. глава 1).
4. Звегинцев, В. А. Мысли о лингвистике / В. А. Звегинцев. – ЛКИ, 2008. – 336 с. – серия «Из лингвистического наследия Звегинцева».
5. Шарафутдинова, Н. С. Теория и история лингвистической науки : учебник / Н. С. Шарафутдинова. – 3-е изд., испр. и доп. – Ульяновск : УлГТУ, 2012. (стр. 6-17)
6. Герд, А. С. Прикладная лингвистика / А. С. Герд. – СПб. : Изд-во СПбу, 2005. – 268 с.

Лекция 2

План:
I. Семиотика. Знак и теория знаков. Язык как языковая система.
II. Информация. Информатика.
III. Компьютерная лингвистика.

I. ***Семиотика*** (греч. *semeoon* – знак) – это наука об общей теории знака, исследующая любые знаковые системы *как средства* обозначения и передачи значения или информации.

Как отдельная наука семиотика оформилась лишь в XX веке, но история этого научного направления довольно интересна. Один из известных современных ученых и писателей Умберто Эко в своей книге «Semiotics and the Philosophy of Language» отмечает, что теоретические концепции о знаках явно просматриваются в работах многих известных миру мыслителей и философов, начиная от Платона и Аристотеля. Умберто Эко также дает широкое определение объекту семиотики как науки: *«Semiotics is concerned with everything that can be taken as a sign»* [1].

Основателем семиотики, как формальной строгой науки, многие считают американского логика и философа Чарльза Пирса, жившего в XIX веке. Пирс определил понятие знака, разделил знаки на классы – индексы, иконы и символы – и установил задачи и рамки новой науки. Идеи Пирса более 100 лет были практически не известны, а его теоретический труд был почти забыт современниками, но помог случай. Имя и труды Пирса получили широкую известность в 30-х годах XX века благодаря другому американскому ученому – Чарльзу Моррису, который адаптировал труды Пирса для широкого читателя, фактически обеспечил им и Пирсу всемирную известность, назвав того основоположником семиотики.

Ч. Моррис в свою очередь возлагал на семиотику большие надежды. Он видел в ней «орудие для объединения всех наук». Поскольку человеческая культура основана на знаковых системах, изучение таких систем является

важной задачей различных гуманитарных дисциплин (психология, логика, философия, лингвистика, культурология и др.). Однако Моррис понимал, что существует весомый недостаток теоретической базы, которая позволила бы обобщить результаты этих дисциплин. Эту проблему Моррис и пытался решить с помощью семиотики как «науки о знаках».

Пирс, Моррис и их последователи развивали логическую линию семиотики, а лингвистическая ветвь семиотики, «изучающая роль знаков в жизни общества», параллельно разрабатывалась в трудах Фердинанда де Соссюра. Швейцарский лингвист Фердинанд де Соссюр также сформулировал основы науки о знаках, назвав ее *семиологией*, и определил ее отношение к лингвистике. Вот что он писал:

«It is ... possible to conceive of a science which studies the role of signs as part of social life. It would form part of social psychology, and hence of general psychology. We shall call it semiology (from the Greek semeoon, 'sign'). It would investigate the nature of signs and the laws governing them. ... Linguistics is only one branch of this general science. The laws which semiology will discover will be laws applicable in linguistics…» [1].

В 60 – 80-е годы XX века вклад в развитие семиотики внесли ученые из СССР. К началу 1960-х годов в Москве сформировалась группа лингвистов (А. А. Зализняк, И. И. Ревзин, В. В. Иванов и др.), пришедших к семиотике разными путями: от структурной лингвистики и машинного перевода, компаративистики и общего языкознания. Параллельно в г. Тарту в области семиотики работал всемирно известный ученый Юрий Лотман. Именно эти ученые стали идеологами той семиотической ветви, которая впоследствии получила название *Московско-Тартуская семиотическая школа*.

Современная международная ассоциация ученых, работающих в области семиотики, *The International Association for Semiotic Studies*, была образована в 1969 году. С этого времени и начинается история современной семиотики как сложной многоплановой науки, которая представляет собой довольно развитую теорию, методы которой позволяют анализировать самые разнообразные сферы

человеческой деятельности. Современная семиотика имеет множество известных продуктивных направлений развития, например:

- семиотика культуры;
- семиотика литературы;
- политическая семиотика;
- семиотика массовых коммуникаций;
- педагогическая семиотика;
- семиотика текста;
- компьютерная семиотика

и другие интересные направления и приложения (см. пример Приложения 4).

Центральным понятием семиотики является понятие знака.

Понятие знака

Определить понятие «знак» пытались многие ученые. Основоположник семиотики Ч. Пирс в своих работах дает 76 определений знака [2]. Мы выберем следующее определение, которое является самым известным и цитируемым определением знака по Пирсу:

*1. «Namely, a sign is something, **A**, which brings something, **B**, its interpretant sign determined or created by it, into the same sort of correspondence with something, C, its object, as that in which itself stands to **C**».*

Это определение в переводе на русский получило свою интерпретацию – *«Знак ... это что-то, что стоит для кого-то вместо чего-то другого в том или ином плане или отношении»* [3].

Также для первоначального объяснения этого термина среди многочисленных определений знака в учебной лингвистической литературе выберем перекликающиеся с определением Пирса.

*2. «Знак – это некое **A**, преднамеренно поставленное кем-то вместо некоторого **C** с целью информировать кого-то об этом **C**».*

Например, значок копирайта © на книге или сайте поставлен, чтобы сказать о защите прав того, кто написал и издал эту книгу или разработал сайт.

Знак материален и обладает направленным значением. В знаке выделяют два основных аспекта (или «плана»):

- *план выражения* (материальный, как выглядит);
- *план содержания* (что обозначает, смысл).

Классическое разделение знаков по Пирсу включает три класса:

1) иконические знаки,

2) знаки-индексы,

3) знаки-символы.

С точки зрения Пирса, **иконический знак** – это знак, который обладает рядом свойств, присущих обозначаемому им объекту. Отношения между знаком и объектом – это отношения подобия; знак оказывается знаком просто в силу того, что ему *«случилось быть похожим»* на свой объект (например, чертежи, схемы, диаграммы, компьютерные иконки в меню редактора и др.).

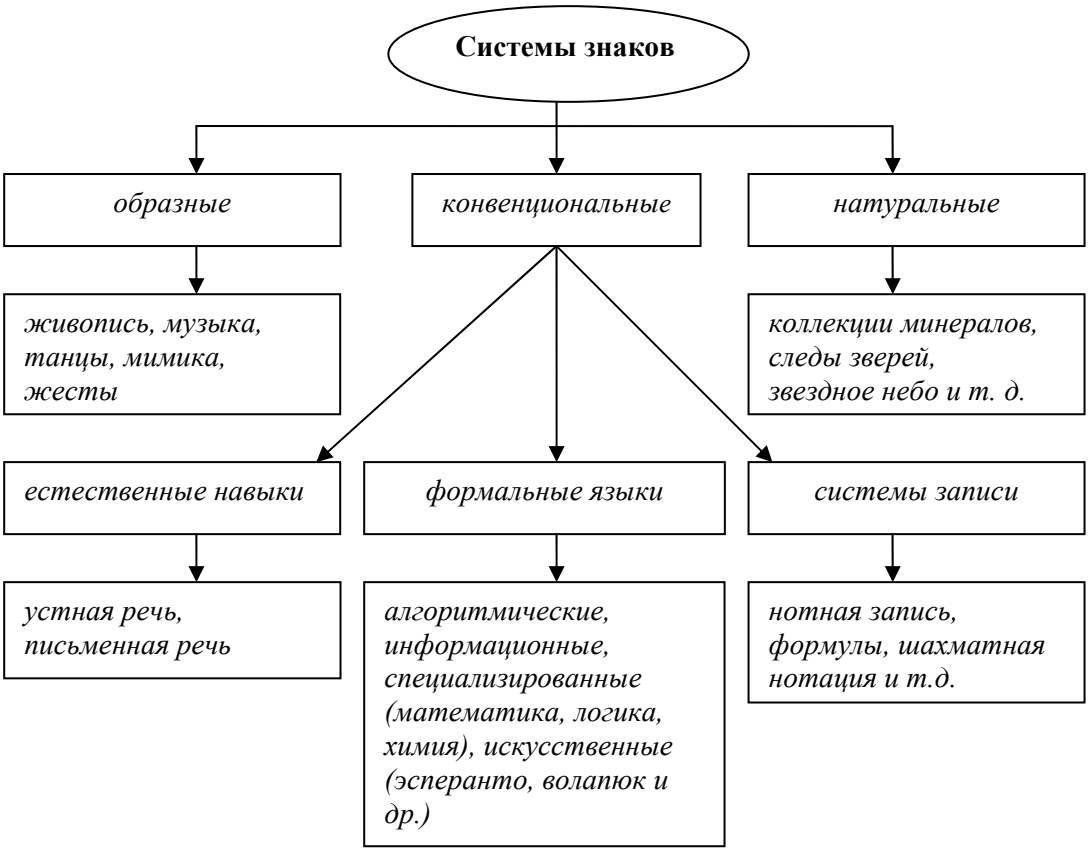
Знаки-индексы – указательные знаки, в которых означаемое и означающее связаны меж собой причинно-следственными отношениями. Пирс приводит такой любопытный пример этого типа знака: *«Таков, например, какой-нибудь предмет с дырой от пули в качестве знака выстрела; ибо без выстрела не было бы никакой дыры, но сама дыра там есть, неважно, достанет ли у кого-нибудь умственных способностей связать ее с выстрелом или нет»* [3].

Знаки-символы – знаки, в которых означаемое и означающее связаны меж собой в рамках некоторой конвенции, то есть по предварительной договоренности. Национальные языки, денежные системы, дорожные знаки – примеры таких конвенций.

Юрий Лотман в свою очередь утверждал, что знаки делятся на условные (символьные) и изобразительные (иконические):

- *Условный* – знак, в котором связь между выражением и содержанием установлена по соглашению группы людей, «внутренне не мотивирована». Самый распространенный условный знак в ЕЯ – это *слово*.
- *Изобразительный* – знак, в котором значение имеет естественно ему присущее выражение. Самый распространенный изобразительный знак – *рисунок*.

Знак – обычно это член определенной знаковой системы.



Например, всем известны такие знаковые системы как дорожные знаки, сигнальные коды (азбука Морзе), чертежные обозначения, компьютерные пиктограммы и т. д.

Лингвистику интересуют общие положения семиотики о знаках, различительные признаки знаков, способы классификации знаков, комбинация их в систему для решения конкретных задач.

Определяя язык, часто говорят, что «Язык – это система знаков». Здесь как раз и просматривается четкая связь лингвистики и семиотики.

Знаковыми единицами языка считают морфемы и лексемы, т. к. они несут информацию и значение в отличие от единиц низшего уровня языка – звуков и букв (так называемые знаки 1-го рода).

Лингвистика также выделяет 3 аспекта знака:

1. Синтактика (изучает отношения между знаками).
2. Семантика (изучает отношения между знаком и значимым).
3. Прагматика (изучает отношения между знаками и теми, кто их использует).

Понимая, что язык – это некоторая система знаков, и говоря, что язык – это *средство* коммуникации (основное определение в прикладной лингвистике), или средство передачи *значения* (определение семиотики), мы находим еще одну связь лингвистики и семиотики. Кроме того, отсюда следует, что язык есть средство передачи *информации*, тем самым мы устанавливаем связь с такой перспективной наукой, как информатика.

В следующем параграфе мы рассмотрим эту взаимосвязь наиболее детально.

II. Первоначально дадим несколько определений понятию *информации* [4].

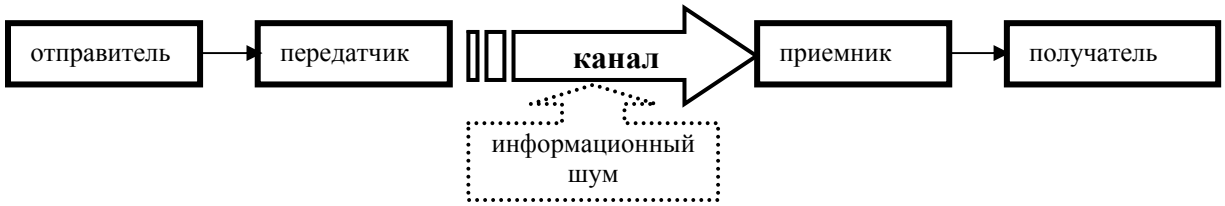
- Под *информацией* понимаются не сами предметы и процессы, а их значения и характеристики (их отображения в виде чисел, текстов, формул, таблиц, графики), т. е. говорят о текстовой, графической информации.

Интересны и такие определения:

- *Информация* есть средство устранения неопределенности в знаниях.
- *Информация* – сведения, данные, несущие значения для конкретного человека.

Вопросами информации занимается наука информатика. Классическая теория информации была обоснована К. Шенноном в 50-х гг. XX века. Основной проблематикой теории информации является оптимизация процессов

передачи информации по каналам связи от отправителя до получателя. Эти вопросы стали наиболее актуальны в наше время с развитием всевозможных средств связи, в т. ч. мобильных и компьютерных.



Информация доступна:

1) восприятию (распознавание образов, текстов и т. д.);
2) передаче (по каналам связи) в виде сигналов (аналоговых/цифровых);
3) обработке (преобразованию в удобную форму);
4) представлению (для демонстрации человеку).

В середине XX века в связи с возрастающими объемами информации и с появлением новых информационных технологий и сред сложилась и стала активно развиваться такая наука как информатика.

Под ***информатикой*** *(Computer Science)* понимают науку о закономерностях записи, хранения, переработки, передачи и использования информации с помощью технических средств.

Вот что также пишет об информатике Захаров В. П.: *«...**информатика** как наука изучает закономерности построения и функционирования автоматизированных систем информационного обмена в различных сферах человеческой деятельности. ... Информатика, кроме своих собственных законов и методов, широко использует методы и результаты других научных дисциплин, в первую очередь, ... теории информации, теории вероятностей, программирования, теории передачи данных, документалистики, библиографии и библиотековедения. Одной из научных дисциплин, тесно связанных с информатикой, является лингвистика. ... И связь эта двухсторонняя: недаром существуют и преподаются такие курсы, как "Лингвистические основы информатики" и "Информационные технологии в лингвистике"»* [5].

Основные понятия в информатике – это *код*, система условных знаков или символов; *алфавит* – набор знаков кода, *текст* – последовательность знаков данного сообщения. Термин «*кодирование информации»* – это представление сообщения в форме, удобной для передачи по данному каналу.

Большинство кодов для кодирования информации основано на тех или иных системах счисления. Передача и хранение информации при этом – это работа с числами. При сравнении систем счисления и реализации компьютерной техники эффективной оказалась двоичная система, т. к. логические элементы должны иметь два устойчивых состояния (включено/выключено или 1;0). С технической точки зрения, *информация* – это сигнал (1,0) или элементарное событие «да» – «нет», измеряемое в двоичных единицах – битах (мера Хартли).

Однако числа, хотя и удобны для передачи и хранения в компьютере, не удобны для восприятия и представления информации для нужд человека. Универсальным же хранилищем и средством передачи информации является естественный человеческий язык. Кроме того, преобладающая часть информации существует в виде письменных и устных текстов на естественном языке.

Как говорит профессор СПбГУ А. С. Герд, *«В целом прикладные аспекты лингвистического обеспечения разнообразных сфер человеческой деятельности сводятся прежде всего к одной общей проблеме – проблеме обработки информации, функционирующей в обществе. Это и текстовая информация в ее письменном виде, и устная речь как наиболее привычный способ коммуникации. Особая роль языкознания в решении практических проблем и потребностей общества определяется самой сущностью естественного человеческого языка, являющегося уникальным средством хранения и передачи информации»* [6].

III. Современный человек живет в глобальной информационной среде, а объемы информации возрастают на порядки ежегодно. Современные

информационные технологии включают растущее число сетевых коммуникационных технологий, автоматизированных информационных систем, средств массовой коммуникации, систем информационного поиска, систем машинного перевода, а техническая коммуникация развивается в сторону оптимизации информационных процессов и использования удобного для человека языка. Проблемы языковой коммуникации «человек – компьютер – человек» и моделирования языка лежат в области исследований такой молодой науки как ***компьютерная лингвистика*** (КЛ, *Computational Linguistics*), которая образовалась на стыке информатики и лингвистики с началом внедрения первых вычислительных машин, а первые эксперименты в этой сфере были связаны с машинным переводом в начале 50-х гг. XX века.

Данное направление стало активно разрабатываться в 60–70 гг., и под ним в первую очередь понималось использование статических методов в языкознании, отсюда и название *«Computational Linguistics»* (в букв. перев. «вычислительная лингвистика»). В России родственные термины «вычислительная лингвистика», «математическая лингвистика» получили распространение в 70-х годах XX века. В конце XX века в связи с развитием компьютерных технологий и их активным приложениям в лингвистических задачах, этот термин как название науки трансформировался, и наука получила более четкое наименование «компьютерная лингвистика».

Большую организационную и научную работу проводит международный орган – Ассоциация компьютерной лингвистики, которая имеет региональные структуры в нескольких странах мира. На официальном сайте этой организации дается общее определение этому прикладному направлению: *«…computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena»* [7].

С точки зрения современного подхода основным направлением компьютерной лингвистики является *Natural Language Processing*, NLP, или

задачи АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЯЗЫКА, что включает *задачи анализа и моделирования языковой структуры*, а именно:

- графематический/фонематический анализ языка;
- морфологический анализ;
- лексико-грамматический анализ языка;
- синтаксический анализ, или парсинг;
- анализ и моделирование семантической структуры;
- задача синтеза языковых элементов, в т. ч. генерация текстов.

Лингвисты также включают следующие ПРИКЛАДНЫЕ НАПРАВЛЕНИЯ в область интересов компьютерной лингвистики. Мы видим, что большинство нижеследующих задач связано с проблемами *автоматической обработки текста*:

- машинный перевод;
- распознавание и синтез речи;
- лингвистические основы информационного поиска;
- автоматическое индексирование, реферирование и классификация текстов;
- автоматический контент-анализ текста;
- авторизация текстов;
- сетевые технологии представления текста и информации на ЕЯ;
- корпусная лингвистика.

Также к задачам компьютерной лингвистики мы относим:

- разработку и использование искусственных языков, в том числе языков программирования, языков информационных систем;
- компьютерную лексикографию и терминографию;
- компьютерную лингводидактику.

Отдельно в рамках компьютерной лингвистики обычно выделяют задачи и технологии использования статистической лингвистической информации, или автоматическую лингвостатистику.

С точки зрения российского восприятия рассматриваемой проблемной области, основную работу в этом направлении проводит российская ассоциация компьютерной лингвистики КОЛИНТ, на сайте которой можно ознакомиться

со всеми научными докладами, представленными на конференции по проблемам компьютерной лингвистики (см. http://www.dialog-21.ru).

Таким образом, ***Компьютерная лингвистика (Computational Linguistics),*** будучи одним из направлений прикладной лингвистики, изучает лингвистические основы информатики и все аспекты связи языка и мышления, моделирования языка и мышления в компьютерной среде с помощью компьютерных программ, а ее интересы лежат в области:

- оптимизации коммуникации на основе лингвистических знаний;
- создания естественно-языкового интерфейса и технологий понимания языка для общения человека с машиной (это одна из основных проблем Искусственного Интеллекта);
- создания и моделирования информационных компьютерных систем.

Семинар

План семинара

I. Понятие знака у Ф.Соссюра. Семантический треугольник (Фреге, Огден и Ричардс).

II. Информация – понимание в разных науках. Примеры лингвистических задач в информатике.

Домашнее задание

- *Пояснить классическое определение Чарльза Сандерса Пирса* «Namely, a sign is something, **A**, which brings something, **B**, its interpretant sign determined or created by it, into the same sort of correspondence with something, C, its object, as that in which itself stands to **C**.», *определив, что представляет из себя **B**, т.е. «interpretant sign».*
- *Информация – в открытых источниках выбрать 3 определения термина, отличающиеся от указанных в лекции.*
- *Охарактеризовать язык математики как знаковую систему.*
- *На сайте КОЛИНТ выписать, какие направления прикладной лингвистики обсуждаются более детально (см. http://forum.dialog-21.ru/actualforum.aspx)*

Список литературы

1. Daniel Chandler. Semiotics for beginners (раздел Introduction) – http://www.aber.ac.uk/media/Documents/S4B/sem01.html
2. Robert Marty. 76 Definitions of The Sign by C. S. Peirce – http://www.cspeirce.com/rsources/76defs/76defs.htm
3. Пирс Ч. Прагматизм. Научная метафизика. http://filosof.historic.ru/books/item/f00/s00/z0000699/st007.shtml
4. Соснина, Е. П. Введение в прикладную лингвистику : учебное пособие / Е. П. Соснина. – Ульяновск : УлГТУ, 2000. – 46 с.
5. Захаров, В. П. Информационные системы (документальный поиск) : учебное пособие / В. П. Захаров. – СПб. : Изд-во СПбГУ, 2002. – 188 с. – http://vp-zakharov.narod.ru/publications.htm
6. Герд, А. С. Предмет и основные направления прикладной лингвистки. – http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html
7. The Association for Computational Linguistics http://www.aclweb.org/archive/misc/what.html

Лекция 3

План:
I. Искусственные языки как знаковые системы.
II. Языки программирования как искусственные языки.
III.Формальные методы описания искусственных языков. Формальная грамматика. Понятие метаязыка.
IV. БНФ-нотации. Синтаксические диаграммы.

I. В предыдущей лекции мы говорили о лингвистических основах информатики и о том, что в связи с невозможностью полной формализации ЕЯ для коммуникации в компьютерных средах необходима разработка специальных искусственных языков (ИЯ), что входит в сферу интересов лингвистики.

Искусственные языки – это знаковые системы, создаваемые для использования в тех областях науки и техники, где применение естественного языка ограничено, менее эффективно или невозможно [1].

Любой искусственный язык по сравнению с естественным всегда ограничен (по словарю, синтаксису, семантике слов) и служит для решения определенных задач.

***Классификация искусственных языков*:**

1. Неспециализированные языки общего назначения (например, эсперанто, волапюк);

2. Специализированные языки различного назначения (например, символические языки наук (математика, логика, химия, физика)).

Во 2-й класс входят *языки человеко-машинного (компьютерного) общения и реализации компьютерных и информационных технологий* (языки программирования, языки операционных систем, языки информационных систем и т. п.).

Лингвист должен иметь представление о том, какие бывают языки, какова структурная организация языков и как создаются искусственные языки. Этому и посвящены следующие наши рассуждения.

II. Языки программирования (ЯП) – это класс искусственных языков, предназначенных для обработки информации с помощью компьютера. Любой язык программирования – это строгая (формальная) знаковая система, при помощи которой записываются компьютерные программы. По разным оценкам, в настоящее время существует от тысячи до десяти тысяч различных языков программирования [2].

Исторически языки программирования возникли в 40-х годах XX в. и качественно совершенствовались в сторону упрощения своего описания, т. е. высокоуровневой организации, методов программирования и приложения для обработки информации человеком.

Можно выделить следующие качественные уровни развития ЯП (т. е. то, как их классифицируют в программировании):

- *Низкий уровень* (работа с машинными кодами, например, есть языки – ассемблеры – это версии машинных кодов, адаптированных под аппаратные платформы компьютеров).
- *Средний уровень*.
- *Высокий уровень* (язык программирования высокого ровня – это язык, команды и структура которого удобны для восприятия человеком, например, Паскаль, Delphi, SQL, Java Script, PHP и др.)

Если взглянуть на язык программирования как на некий объект с лингвистической точки зрения, можно увидеть, что ЯП так же, как и любой язык, имеет *свои ярусы* или *структуру*.

Первый низший уровень – *символьный*, его элементы алфавита – буквы, спецсимволы (по аналогии с ЕЯ – графематический уровень).

Второй уровень – это *уровень имен*, например, зарезервированных слов, выражений (в ЕЯ – это лексический уровень).

Третий уровень – *операторный* (командный), где синтаксические конструкции имеют повелительный характер (в ЕЯ – аналог синтаксического уровня), и последний – уровень *программы*, всегда являющейся синтаксически

и семантически законченной последовательностью предписаний-команд. Программа – это структурно строгий текст, записанный по формально заданным правилам искусственного языка программирования (см. пример текстов программ в Приложении 5).

III. Общим признаком описания специализированных искусственных языков является ***формальный метод*** их описания и определения путем задания *алфавита, словаря* и *системы правил образования и преобразования выражений* (грамматика). Формальный метод служит для порождения «*правильных выражений*» («правильных» – значит «записанных по определенным правилам»).

Например, для языков программирования задаются определенные формы языковых элементов (алфавит, слова), правила построения команд, текстов программ, т. е. точно описываются семантика и синтаксис для однозначного понимания программ компьютером.

Вообще, при написании «правильных выражений», т. е. для формального описания синтаксиса элементов любого языка, широко используются такие нотации как формальные грамматики.

Формальная грамматика – это система строгих (часто математических) правил, позволяющая с помощью единообразных процедур получать (выводить) правильные выражения данного языка либо анализировать имеющиеся выражения на предмет их соответствия правилам языка.

Вопросами формальных грамматик и теорией формальных языков занимается такой раздел языкознания как ***математическая лингвистика***. Она является смежным направлением прикладной лингвистики, тесно соприкасающимся с математикой и информатикой.

В 1957 году в своей книге «Syntactic structures» американский ученый-лингвист Ноам Хомский предложил классификацию формальных языков по типу правил формальной грамматики [3].

Существует множество видов формальных грамматик, как например:

1. Регулярная грамматика.
2. Контекстно-свободная грамматика.
3. Грамматика непосредственно-составляющих.
4. Лексико-функциональная грамматика.
5. Грамматика Монтегю.

Кратко опишем ***порождающую грамматику***.

Порождающая формальная грамматика – это система

$$\mathbf{G = <V_т, V_{нт}, S, R>},$$

где **G** – грамматика;
$\mathbf{V_т}$ – множество терминальных (конечных) символов языка;
$\mathbf{V_{нт}}$ – множество нетерминальных символов (из которых можно выводить далее), заключаются нами в примере ниже в угловые скобки <…>;
S – начальный символ нетерминального множества;
R – система правил вывода типа X→Y (где X,Y – цепочки символов из $V_т$, $V_{нт}$).

Множество цепочек, выводимых через **G** из ее начального символа S, есть выражения языка, порождаемые этой грамматикой **G** (т. е. вывод цепочек всегда начинается с нетерминала S).

Пример:

Формальная система:	Система правил R:	Выражения, порождаемые согласно синтаксическим правилам R:
Г =<{I, We, They, like, music .}, {S, Pr, V, N}, S, R>,	**<S>→<Pr><V><N>.**	
	<Pr> → I | We | They	**I like music.**
где {I, We, They, .} – $V_т$,	**<V> → like**	**We like music.**
{S, Pr, V, N} - $V_{нт}$	**<N> → music**	**They like music.**

Формальная грамматика, изложенная по подобным правилам, в свою очередь, работает на базе ***метаязыка***, т. е. специальной вспомогательной системы знаков (нотации) для работы с конечным языком.

IV. На практике применяется еще один метаязык, который даже считают синонимичной записью или функциональным аналогом нотации формальных грамматик. Это ***Бэкус-Науровы формы*** (БНФ–формы или

БНФ–нотации), которые, как и формальная грамматика, служат для задания правил получения правильных выражений и текстов.

Пример:

Пример типовой странички словаря

P

Pay [pei] 1. платить;
2. заработная плата;
3. расплата.

Pea [pi:] горох.

Peak [pi:k] остроконечная вершина.

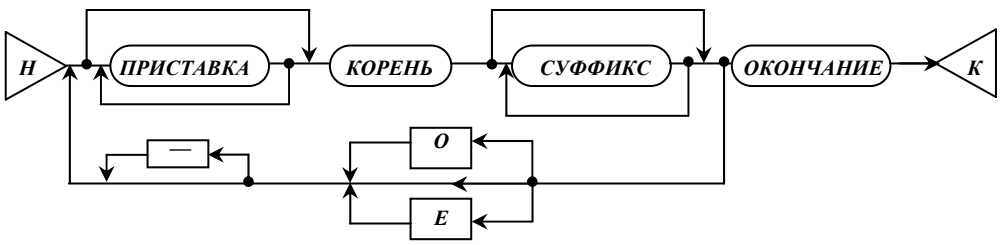
БНФ-нотация для описания англо-русского словаря:

<словарь> ::= [<раздел>]
<раздел> ::= <заглавная лат.буква>[<словарная статья>]
<заглавная лат.буква> ::= A | B | …. | Z
<словарная статья> ::= <термин> <транскрипция> <перевод>.
<термин> ::= <заглавная лат.буква> [<прописная лат.буква>]
<прописная лат.буква>] ::= a | b | … | z
<транскрипция> ::= [[<фонетический знак>]]
< фонетический знак > ::= a: | э | … | z
<перевод> ::= <определение 1> | <определение 2>
<определение 1> ::= <слово> | <словосочетание>
<слово> ::=[<прописная русск.буква>]
<прописная русск.буква>::= а | б | в…. | я
<словосочетание> ::= [<слово> _]
<определение 2> ::= **1.**<определение 1>;
2.<определение 1>;
3.<определение 1>

Аналогичный метаязык, имеющий графическое наглядное представление – это ***синтаксические диаграммы,*** которые используются часто при преподавании программирования на языках высокого уровня. Синтаксическая диаграмма – это схема, объясняющая правило построения либо некоторого элемента, выражения, либо текста.

Пример:

Синтаксическая диаграмма морфологической структуры русского слова*:



* Диаграмма представлена нами в ограниченном виде.

Обе эти формы нотаций нашли широкое применение при описании языков программирования в информатике. Для прикладной лингвистики построение формальных грамматик, БНФ и синтаксических диаграмм интересно как способ понять структуру любого языка, увидеть возможности по моделированию искусственных языков и лингвистических структур.

Семинар

План семинара

I. Линейные и ветвящиеся синтаксические диаграммы.

II. Пример БНФ-нотации.

III. Примеры формальных грамматик и вывода на них.

Домашнее задание

- *Составить свой пример формальной грамматики.*
- *Составить БНФ-нотацию для описания структуры текста программы телевидения.*
- *Решить синтаксическую задачу №1 Приложения 3.*

Список литературы

1. Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. – М., 1998. – С. 201-207, 287-289, 615.
2. Свободная энциклопедия языков программирования - http://progopedia.ru/
3. Chomsky, Noam. Syntactic structures. Walter de Gruyter, 2002. – 117 с. - books.google.com

Лекция 4

План:

I. Моделирование и модель в лингвистике.

II. Модели знаний в искусственном интеллекте. Языки представления знаний как варианты искусственных языков.

I. В разных науках приходится иметь дело с различными моделями (образцами) тех или иных процессов или объектов исследования этих наук. ***Метод моделирования*** – центральный исследовательский метод практически в любой науке, нацеленный на выяснение либо способов, правильности функционирования объекта, либо его свойств при помощи построения модели этого объекта. Метод моделирования языка и языковых процессов широко используется лингвистами, т. к. дает возможность реализовать теоретические знания на практике. Прикладная лингвистика стремится строить модели, отображающие конкретные лингвистические объекты, их системы, а также процессы речемыслительной деятельности человека.

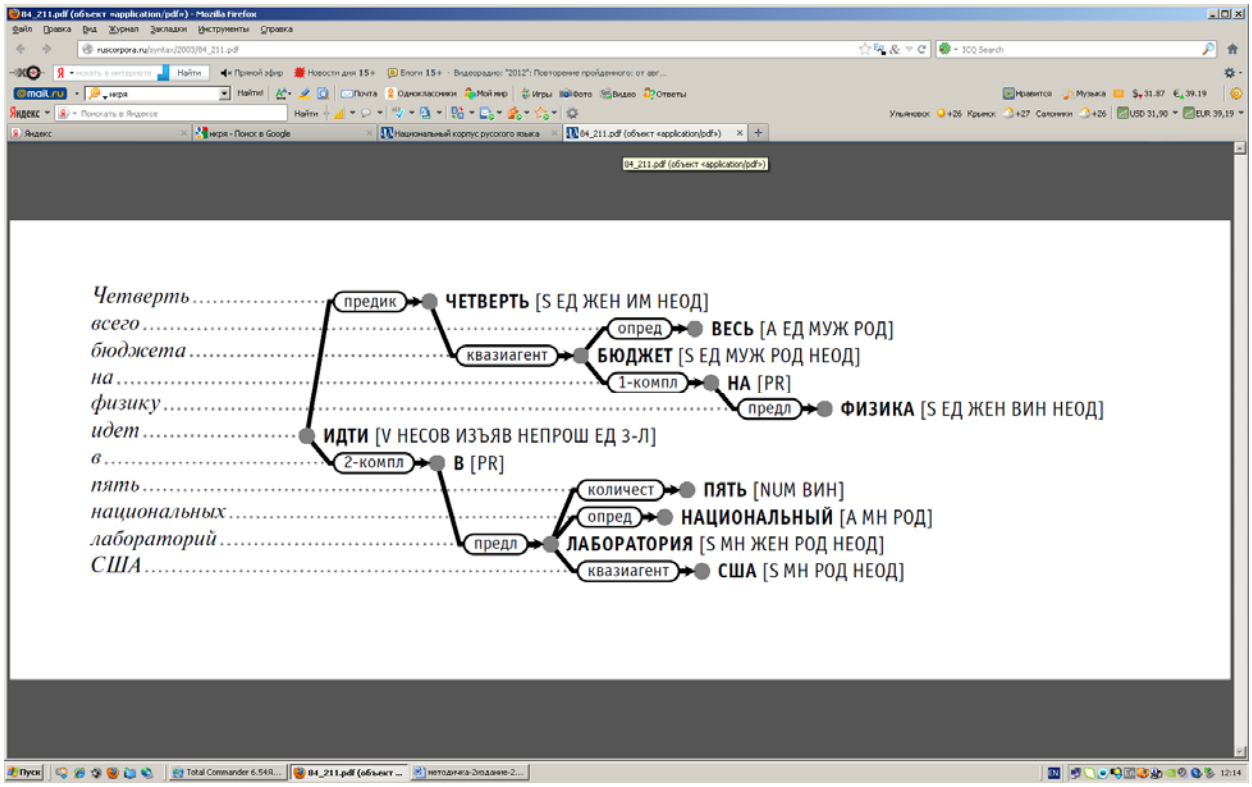
Понятие лингвистической модели также активно используется в компьютерной лингвистике, т. к. без построения лингвистических моделей невозможно решить многие практические задачи создания и использования лингвистических ресурсов (например, машинных переводчиков, электронных словарей) и задачи автоматической обработки языка.

Всякая модель строится на основе гипотезы о возможном устройстве оригинала и зачастую представляет собой функциональный аналог оригинала, что позволяет переносить полученные знания с модели на оригинал. Критерием адекватности модели часто является эксперимент.

Даже в лингвистике, не говоря о других науках, существует множество определений понятия **«модель»**. Чаще всего под моделью понимают формальный способ представления, описания, лингвистических объектов, например:

- *модель–схемы/описания* для удобного представления структуры и содержания языковых объектов (модель словарной статьи словаря или см., например, синтаксическую диаграмму ниже).

Пример синтаксической диаграммы – дерева зависимостей*



* Пример на рисунке построен на базе Национального корпуса русского языка

- *модель-тип (language pattern)* каких-либо устойчивых типовых единиц исследуемого языка (сочетаемостей слов, синтаксических структур).

Такие устойчивые образцы мы неоднократно встречаем в грамматиках при изучении иностранных языков. Например, это стандартная для английского языка структура предложения, которую называют **SVO** или **Subject-Verb-Object pattern** (пр. *The Big Brother is watching you*).

- *модель-теория* – описание на базе строгой формализованной научной теории, т. е. структура элементов и правил с фиксированным метаязыком (например, теория формальных грамматик).

Таким образом, ***модель*** в лингвистике – это формализованная структура или описание с фиксированным метаязыком, служащая образцом для исследования порождаемого языка, анализа его характеристик и функций. Модель всегда предполагает наличие однозначно заданных объектов, связывающих их отношений и правил обращения с ними (см. пример формальной грамматики в лекции 3).

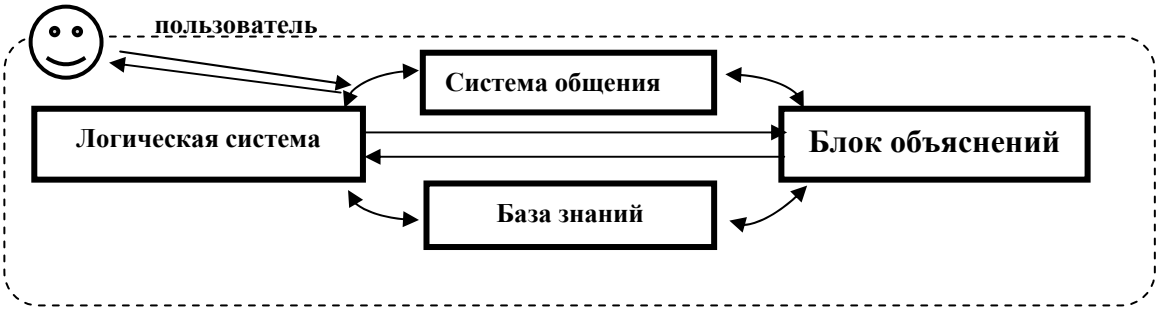
Понятие лингвистической модели возникло в структурной лингвистике, но вошло в активный научный обиход в 60–70 гг. XX в. с развитием математической лингвистики и проникновением в лингвистику точных формальных методов исследования. Лингвисты обратили особое внимание на семантику единиц языка и на модели речевой деятельности, тесно связанные с моделями мышления (которые часто исследуются в логике). В 70-х годах XX в. в результате неудачных попыток всестороннего полного моделирования естественных языков в рамках решения задач машинного перевода *«ученые пришли к выводу, что решение многих прикладных проблем не может быть чисто лингвистическим, а лежит на совсем иных путях, на путях моделирования поведения и мышления человека, семантики, синтеза формальных и семантических средств языка. Так появилась одна из важнейших межотраслевых фундаментальных проблем прикладной направленности – проблема моделирования знаний»* [1].

II. Моделирование на базе компьютерных технологий человеческих знаний, понимания языка и человеческого мышления – это задачи ***Искусственного интеллекта (Artificial Intelligence)*** как одного из ведущих научных направлений компьютерной науки – информатики, которое занимается созданием интеллектуальных когнитивных технологий и машин, способных понимать, моделировать и анализировать тексты, хранить и перерабатывать естественно-языковую информацию, принимать интеллектуальные решения.

Система «искусственного интеллекта» должна понимать вопросы человека, решать интеллектуальные задачи и вести диалог с человеком на

основе заложенных в нее процедурных и декларативных знаний. Примером интеллектуальной искусственной системы является ***экспертная система***, качество которой определяется в первую очередь тем, насколько естественно общение с ней человека при решении задач.

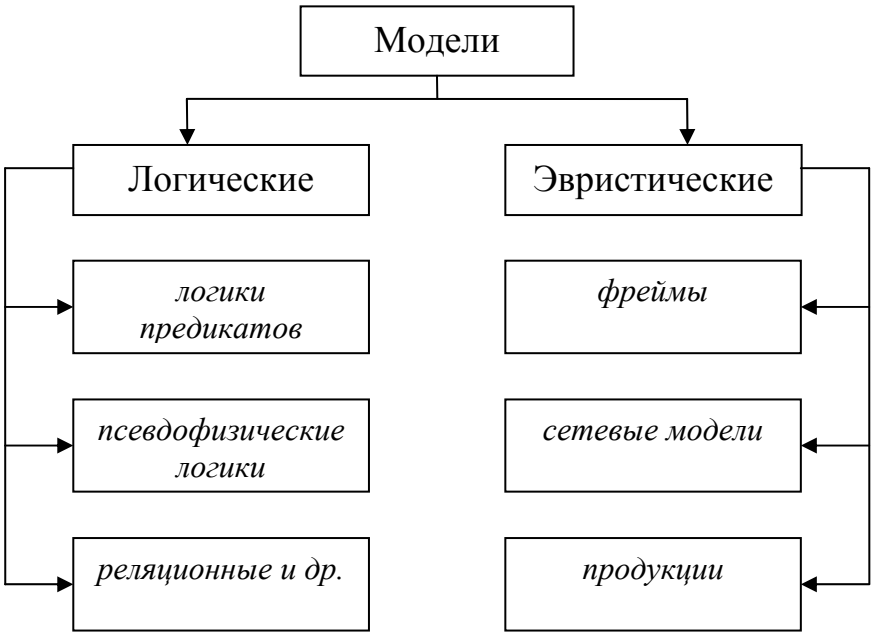
Структура экспертной системы



В искусственном интеллекте важно понятие ***«знания»***, а при построении интеллектуальных систем – *языки представления знаний*.

Знания, как правило, включают в себя 3 составляющие – *опыт, навыки* и *умения* – и бывают ***декларативными*** (*«знаю ЧТО это»)* и ***процедурными*** *(«знаю КАК это делать»*).

Знания в интеллектуальной системе представляют специальными моделями на искусственных языках лингвистического обеспечения этой системы.



Для прикладной лингвистики интересны ***языки представления знаний и семантики*** (ЯПЗ). ЯПЗ – это искусственные языки, построенные по законам искусственных языков. Наиболее популярные ЯПЗ – это логические, сетевые, фреймовые и продукционные [2, 3].

а) ***Логические ЯПЗ*** представляют знания в виде синтаксически правильных формул какой-либо формальной логической системы.

Логических систем очень много. Разработаны даже псевдофизические логики, т. е. системы правил, описывающих отношения реального мира, например, логики времени, причины и следствия, логики пространственных отношений [4]. Часто используют ***логику предикатов*** и ее язык логических формул.

Примеры логических формул:

1. ***Вася любит Машу***.

Это предложение на ЕЯ можно записать по-другому

Любить (Вася, Маша).

или на языке логики предикатов.

L(b, m)

Предикатная формула может быть одноместная и многоместная.

2. ***Вася не студент.*** ***Вася подарил Маше кольцо.***

~ S (b) **P (b, m, k)**

Часто предикатная формула может быть вложенной

3. Теорема Пифагора на языке логической формулы будет иметь вид:

РАВНЫ [СУММА(КВАДРАТ (1-Й КАТЕТ), КВАДРАТ(2-Й КАТЕТ)), КВАДРАТ(ГИПОТЕНУЗА)]

или на символьном метаязыке

P1(P2(P3(k1), P3(k2)), P3(g))

4. В логических формулах часто используются специальные знаки – универсальные кванторы, например, ∃ (существует…), ∀ (для всех…).

Есть девушки красивее Маши.	***Клиент всегда прав.***
∃ x (G(x) & N(x, m))	**∀x (C(x) →P(x))**

Попробуйте подобрать свои примеры про любовь (обозначим ее как **L**) на естественном языке для таких формальных записей:

∃ x L(x,x)	**∀x (L(x,m) →L(x,d))**
~ ∃x ∀y L(y,x)	**~ ∀x∃ y (D(y) & L(x,y)**

А теперь попробуйте преобразовать следующие два предложения в формальную запись:

Не все студенты ответят на все вопросы.

Вы сможете, если попытаетесь.

Решение задач в логике на логическом ЯПЗ – это логический (дедуктивный) вывод по правилам этой логической системы.

б) ***Сетевые ЯПЗ*** в качестве моделей используют *семантические сети*, где узлы сети – это какие-либо информационные единицы – понятия, факты, процессы, имена, а дуги – отношения между ними. Отношения могут быть любыми (временные, причинно-следственные, больше-меньше и т. п.).

Сетевые ЯПЗ часто используют для явного описания отношений в той или иной ситуации или для описания семантики структур.

Решение задач на сетевых моделях сводится к поиску фрагмента сети, совпадающему с данным образцом и к организации логического вывода на семантической сети. Связи элементов семантической сети можно представить в виде формул – записей на метаязыке семантических сетей, т.к. графически можно представить наглядно лишь простые структуры.

Пример:

Дано такое предложение на ЕЯ:

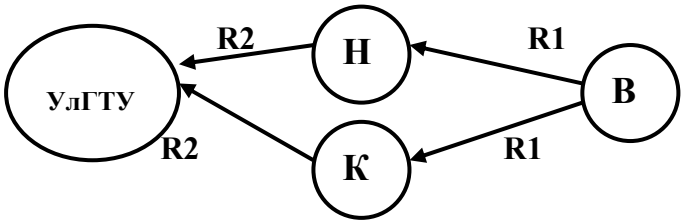
«Вася видел Наташу и Катю в УлГТУ».

Вариант семантической сети

для этой ситуации будет:

R1 – отношение «видеть»,

R2 – отношение «находиться в».



Решение задачи на сети

Если у компьютерной интеллектуальной системы с таким знанием ситуации спросить *«Вася видел Катю?»* (т. е. есть ли связь (**В R1 К**)), то она ответит утвердительно, т. к. такая связь в сети действительно присутствует. Если же спросить *«Катя видела Васю?»* (**К R1 В**), то такая связь в сети не указана, и система ответит отрицательно.

в) ***Фреймовые ЯПЗ***

Фрейм (*frame* – каркас, скелет, рамка) – это шаблон типовой ситуации или некоторая структура, содержащая сведения об определенном объекте, его характеристиках или их значениях и выступающая как целостная единица знаний. Понятие «фрейма» в классическом понимании связано с минимальным описанием факта или явления, у которого нельзя удалить никакую часть описания без того, чтобы не утратить полноту этого описания.

Фрейм может быть *ролевой* или *структурный*.

Структурный фрейм в большей мере отражает декларативные знания, т. е. описывает структуру какого-либо понятия, объекта или документа. Представьте структуру Вашего фрейма в социальной сети.

Ролевой фрейм, в отличие от структурного, представляет процедурные знания, например, *скрипты* или *сценарии* каких либо типовых процедур или работ. Скрипты широко используются в программировании бизнес-

приложений и стандартных автоматов, например, при оплате услуг в банкоматах или для дистанционных операций. Одна из более сложных областей использования ролевых фреймов – это робототехника.

Основные характеристики этих ЯПЗ – компактность и вложенностью уровней.

Пример формальной записи сценария и модель варианта сценария:

БНФ-нотация сценария [2]:	***Сценарий посадки в самолет:***
<СЦ>: : = (**СЦЕНАРИЙ** <имя СЦ> : <имя слота1>: <Значение слота 1>; . . . . . . <имя слота n>: <Значение слота n>). <имя СЦ> ::= <слово>|<словосочетание> <имя слота>:: = **деятель** |**участники**| **цель деятеля**|**цель участников** |**место** |**время** |**посылки**|**следствия**|**средства реализации**| **побочные действия** |**ключ** <значение слота>:: = «<текст>» | «<текст>»R_1 «<текст>» . . . . . R_n<текст>	(**СЦЕНАРИЙ** *Посадка в самолет*: **деятель:** «сотрудник авиакомпании»; **участники:** «пассажиры рейса N 186»; **цель деятеля:** «зарегистрировать всех пассажиров» R1 «посадить их в самолет»; **цель участников:** «пройти регистрацию, сдать багаж и сесть в самолет»; **место:** «аэропорт Ульяновск-Восточный»; **время:** «в течение 2 часов до отлета рейса N 186»; **посылки:** «регистрация» R1 «сбор на посадку» R1 «контроль на таможне» R1 «подача автобуса» R1 «подача трапа»; **следствия:** «заполнение самолета пассажирами»; **ключ:** «посадка в самолет всех пассажиров, прошедших регистрацию»).

В примере **ключ** – основное событие или тип сценария, **посылки** – необходимое условие реализации сценария (последовательность действий для ключа), **следствия** – результат выполнения сценария. Сценарий завершается, если реализовано ключевое событие (ключ) и достигнута цель деятеля. R1 – отношение времени «раньше».

г) ***Продукционные ЯПЗ***

Продукции – это одна из популярных форм представления процедурных знаний в экспертных системах и других системах знаний, работающих со сценариями.

Продукции представляют в основном процедурные знания, причем формула правил продукции относительно проста – *«ЕСЛИ ..., ТО ...»:*

if <condition> **then** <conclusion>

или

if <condition> **then** <action>

Например,

ЕСЛИ < температура тела более 40°>,

ТО <срочно *вызывай скорую помощь по телефону 03*>.

Продукции могут быть составными, например,

if a and b and c **then** d.

В форме правил-продукций можно задавать и лингвистические знания о языке, полученные в результате анализа множества текстов.

Задачи для продукционной модели ставятся как задачи поиска нужной последовательности продукций, при котором достигается нужная цель.

Семинар

План семинара

I. Сетевые модели. Примеры семантических сетей и вывода на них.
II. Логики. Псевдофизические логики и их правила.
III. Фрейм как шаблон типовой ситуации. Сценарии.

Домашнее задание

- *Построить семантическую сеть для 5 любых предложений по выбору, записать в виде формул.*
- *По БНФ составить сценарий «тушения пожара».*
- *На базе любой социальной сети разобрать любой фрейм (структурный).*

- *Построить текстовую структуру заданного примера текста, дополнив ее с помощью правил псевдофизической логики.*

Список литературы

1. Герд, А. С. Предмет и основные направления прикладной лингвистки. – http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html

2. Искусственный интеллект // Справочник «Модели и методы». – М., 1990. – Т.2. – С. 7–60.

3. Поспелов, Д. А. Моделирование рассуждений / Д. А. Поспелов. – М., 1989. – С. 124-151.

4. Кандрашина, Е. Ю. Представление знаний о времени и пространстве в интеллектуальных системах / Е. Ю. Кандрашин, Д. А. Поспелов. – М., 1989. – С. 8-34, 248-258.

Лекция 5

План:

I. Информационный поиск.

II. Лингвистика в задачах информационного поиска. Информационно-поисковые языки как искусственные языки.

I. Теория и практика информационного поиска стала развиваться в середине XX века. К концу века с внедрением мировых информационных компьютерных сетей и их информационно-лингвистического обеспечения эта область стала отдельным признанным научно-практическим направлением компьютерной науки со своей теорией информационного поиска и многочисленными практическими разработками, продуктами и технологиями.

Поиск информации (*Information Retrieval*) – это процесс отыскания в некоторой системе хранения информации таких документов (текстов, записей и т. д.), которые соответствуют поступившему запросу. В качестве таких средств хранения и поиска информации выступают информационно-поисковые системы (*IRS/Information Retrieval Systems*), элементами которых являются структурированный массив документов (база данных/индекс), выступающих как объект поиска, разнообразные технические и программные средства, как например, программы-роботы, а также информационно-поисковый язык, задающий правила индексирования документов, правила поиска.

По В. П. Захарову «***Информационно-поисковые система*** (ИПС) – *это упорядоченная совокупность документов (массивов документов) и информационных технологий, предназначенных для хранения и поиска информации – текстов (документов) или данных (фактов). Информационно-поисковыми системами являются любые определенным образом организованные хранилища информации. Причем информационно-поисковые системы могут быть и неавтоматизированными. Главное – это целевая функция: хранение и поиск информации*» [2].

При вводе документа в базу данных ИПС его индексируют (поэтому саму базу поисковой системы часто называют *индексом*). ***Процесс индексирования*** в

основном связан с определением и выборкой ключевых слов обрабатываемых документов и выражением их формально в виде *поискового образа* [3]. Так база данных ИПС состоит из множества индексных поисковых образов. Непосредственно при поиске производится сопоставление вашего запроса, т. е. того, что вы указали в запросе с поисковым образом, т. е. тем, что хранится в индексе.

Для более наглядного понимания способа хранения информации в базе ИПС приведем следующий простой пример.

Пример:

АДРЕСА документов	***Словарь информационно-поискового языка***						
		S1	**S2**	**S3**	**S4**	**S5**	**S6**
	1	♥	♥			♥	
	2			♥	♥		♥
	3	♥	♥	♥			
	4			♥	♥		
	5					♥	♥

Здесь: **S1-S6** – дескрипторы (ключевое слово); ♥ – условный знак того, что данный дескриптор находится в документе;
1-5 – адреса или номера документов.

Например, документ N 3 имеет ПО ={S1, S2, S3}, а документ N5 – {S5, S6}.

Если поисковое предписание-запрос пользователя будет ПП= **S1 AND S2**, то поисковая система выдаст адреса документов N1 и N3.

Если поисковое предписание-запрос пользователя будет ПП=**S5 OR S6**, то поисковая система выдаст адреса документов N1, N2, N5.

В зависимости от объекта хранения и типа запроса часто в учебной литературе различают два вида информационного поиска [2, п.1.1]:

- *документальный* (когда пользователь ищет какой-либо текст/документ);
- *фактографический* (когда пользователь ищет какие-либо фактические данные, например, *«День рождения Анджелины Джоли»*).

Характеристики информационного поиска – это такие его семантические показатели как:

- *полнота выдачи информации;*
- *точность ее выдачи;*
- *потери информации*;
- *информационный шум.*

Полнотой поиска *(Recall)* называется мера, вычисляемая как отношение количества *выданных релевантных* документов к *общему числу релевантных* документов, содержащихся в базе информационно-поисковой системы.

Точность поиска *(Precision)* – это отношение количества *выданных системой релевантных* документов к *общему числу документов в выдаче*.

Рассмотрим следующую таблицу, по которой легко просчитываются основные показатели информационного поиска:

Документы	**релевантные (необходимые, нужные)**	**нерелевантные (ненужные)**
выданные ИПС	**a**	**b**
невыданные ИПС	**c**	**d**

ПВ= (a/(a+c))·100% – полнота выдачи информации

ПИ= (c/(a+c))·100% – потеря информации

ИШ= (b/(a+b))·100% – информационный шум

ТВ= (a/(a+b))·100% – точность выдачи

Особенно интересно нам понятие «*релевантность»* как фундаментальное понятие теории информационного поиска. Согласно определению, документ, центральный предмет или тема которого в целом соответствует смысловому

содержанию информационного запроса, называется *релевантным*, а *свойство смысловой близости* между документом и информационным запросом – *релевантностью*. По многим, чаще всего субъективным, причинам релевантность была и остается основной проблемой информационного поиска.

II. Теория и практика информационного поиска тесно связана с лингвистикой, т. к., во-первых, основной объем информации и тексты документов представлены на естественных языках, а для извлечения информации и индексирования текстов необходимы знания автоматической обработки языков, например, правила компьютерной морфологии. Во-вторых, информационные компьютерные системы построены и работают на базе искусственных языков.

Из определения информационно-поисковой системы видно, что ее основным лингвистическим средством является специализированный искусственный язык.

Информационно-поисковый язык (ИПЯ) – это специализированный искусственный язык, предназначенный для:

1) описания информационных запросов к информационно-поисковым системам (***язык запросов***),

2) описания формальных характеристик документов в виде поискового образа, хранящегося в базе системы ***(язык индексирования)***.

Необходимость внедрения искусственных языков вызвана необходимостью устранения избыточности естественного языка для информационного поиска, а также ликвидации языковой синонимии, омонимии и неоднозначностей разного рода.

Информационно-поисковый язык, как и любой язык, состоит из фиксированных единиц, например, имеет свой словарь и синтаксис, и является искусственным языком, т.е. ограниченным по своей форме и структуре стоящими перед ним задачами на поиск.

Рассмотрим каждый из вариантов ИПЯ более подробно:

2.1. Языки запросов

Кажущиеся простыми на первый взгляд *языки запросов* представляют собой довольно комплексные системы правил и процедур. Как правило, модель языка запросов включает следующие элементы:

- Поисковые единицы (ключевые слова, выражающие информационную потребность пользователя).
- Средства лемматизации лингвистических единиц запроса (приведение слов к нормальной словарной форме).
- Специальные поисковые (булевские) операторы, типа OR, NOT, AND.
- Средства линейной грамматики (операторы расстояния, позиционные операторы, например, title:).
- Дополнительные условия поиска: ограничение области поиска по языку, региону, дате создания документа и т. п.

Запрос на поиск чаще всего ограничен по языку и может иметь определенное формальное представление. Самый частотный способ поискового запроса – через ограниченный набор ключевых слов, задаваемый пользователем в поисковой строке. Например, найти эту книгу в Интернете можно, задав ключевые слова *Соснина введение прикладная лингвистика* в поисковых системах Яндекс, Google и т. п.

Кроме стандартной практики в поисковых системах предусмотрена функция расширенного языка запросов, ее можно легко найти на сайтах поисковых машин – http://yandex.ru/search/advanced или http://www.google.ru/advanced_search. Расширенный поисковый язык включает множество операторов для улучшения качества и сужения зоны поиска релевантной информации.

Например:

- *"чемпионат мира 2014"* ***OR*** *"олимпийские игры 2014";*
- *мумий тролль мультфильм* ***-****рок* ***-****лагутенко* – Поиск в Яндексе или Google только мультфильма, а НЕ (**-**отрицание) группы Ильи Лагутенко;

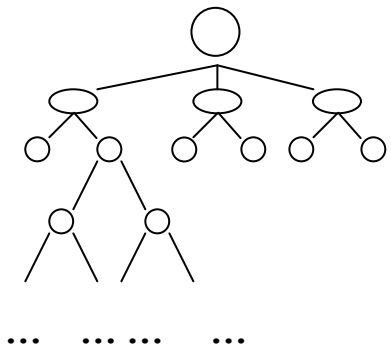
- *социолингвистика* ***-site:****ru.wikipedia.org* (ищем информацию везде кроме сайта русской Википедии);
- ***date:****ГГГГ{*|ММ{*|ДД}}* – Поиск в Яндексе только по страницам, дата которых удовлетворяет заданному условию (например, ищем труды профессора УлГТУ только за 2012 г. *Шарафутдинова Н. С.* ***date: 2012***).

2.2. Языки индексирования

В литературе по теории информационного поиска [2, 4, 5] обычно выделяют следующие распространенные *виды информационно-поисковых языков* для индексирования документов:

1. Языки классификаций – иерархические; алфавитно-предметные и др.;
2. Дескрипторные языки.

Иерархические ИПЯ (иерархия – классификация от общего к частному). Такая организация используется для поиска книг в библиотеке (например, Универсальная десятичная классификация – УДК).



Дерево классов

Алфавитно-предметные ИПЯ представляются как алфавитный список ключевых слов какого-либо документа с пометами (например, алфавитно-предметный указатель в конце книг). Используется для построения различных указателей, каталогов, картотек.

Дескрипторные языки – наиболее естественная и популярная форма языков индексирования для выражения поисковых образов и в настоящее время

широко используется в современных информационно-поисковых системах (в частности во многих поисковых системах сети Internet).

В таких языках используется принцип координатного индексирования, т. е. перечисляются ключевые слова или дескрипторы, которые выражают центральную тему или целостную характеристику искомого объекта. При этом используется принцип логического умножения понятий, в результате которого из простых лексических единиц строятся более сложные, выражающие более узкие понятия. Этот принцип описания содержания документов через перечисление ключевых слов существует издавна. Например, пересечением понятий ПАРИЖ и ДОСТОПРИМЕЧАТЕЛЬНОСТИ, заданных в запросе, порождается новое более узкое понятие ДОСТОПРИМЕЧАТЕЛЬНОСТИ ПАРИЖА.

В качестве лексических единиц в дескрипторных ИПЯ выступают *дескрипторы* – имена понятий или классов понятий, которые явно перечисляются в дескрипторном словаре. Это слова (или словосочетания), выбранные в качестве представителей классов и групп синонимичных слов и словосочетаний. Как правило, это существительные или номинативные выражения.

Списки дескрипторов организованы в специальные семантические словари (поисковые тезаурусы). *Поисковые тезаурусы (одноязычные или многоязычные)* – специальные словари для информационного поиска по массивам естественно-языковых документов, организованные по принципу сопоставления слов с их понятиями. Их структура и разработка часто стандартизируются, см., например, ГОСТ 7.25-2001.

Такой дескрипторный словарь используется как средство лексического контроля (например, снятия омонимии, синонимии единиц) при индексировании документов и запросов. Например, при индексировании все синонимы запроса и поискового образа представляются одной и той же лексической единицей – дескриптором (ср. *лингвистика – языкознание языковедение, наука о языке*).

Структурно поисковый тезаурус представляет собой алфавитный список дескрипторов вместе с их словарными статьями (гнездами). Словарная статья обычно содержит:

- заглавный дескриптор;
- ключевые слова или словосочетания, входящие в гнездо данного дескриптора (условные синонимы);
- «вышестоящие» дескрипторы (находящиеся с данным в отношении «*род–вид», «часть–целое»);*
- «ижестоящие» дескрипторы (находящиеся с заглавным дескриптором в отношении «*вид–род», «целое–часть»*);
- ассоциативные дескрипторы (связанные с данным другими разнообразными отношениями: *причина–следствие, процесс–объект, свойство–носитель свойства, функциональное сходство*).

Указанные отношения приводятся с пометами, чаще всего так:

ДЕСКРИПТОР
с – ключевые слова (синонимы)
в – родовые слова
(дескриптор, подчиняющий данный)
н – видовые слова
(дескриптор, подчиненный данному)
а – ассоциации (отношения)

Рисунок. *Словарная статья информационно-поискового тезауруса*

Нужно отметить, что тезаурусы бывают не только информационно-поисковые, а их структура не ограничивается только представленными отношениями. Тезаурусы относятся к очень интересному классу семантических словарей, что мы рассмотрим в следующей лекции, посвященной одному из древних практических направлений лингвистики – лексикографии.

Семинар

План

I. Хранение информации в ИПС.

II. Тезаурус → дескрипторная статья.

Домашнее задание

• *Составить 5 дескрипторных статей по лекциям для ИПС. Составить матрицу ИПС по лекциям. Составить запросы.*

• *Определить какие еще отношения, кроме перечисленных выше в словарной статье тезауруса, указаны в стандарте [6]. Привести примеры.*

Список литературы

1. Web information retrieval – http://research.microsoft.com/en-us/collaboration/global/asia-pacific/talent/webirhistoryfuturetrends.pdf
2. Захаров, В. П. Информационные системы (документальный поиск): учебное пособие / В. П. Захаров. – СПб. : Изд-во СПбГУ, 2002. – 188 с. – http://vp-zakharov.narod.ru/publications.htm
3. Баранов, А. Н. Введение в прикладную лингвистику /А. Н. Баранов. – М. : Эдиториал УРСС, 2001. – 360 с. (стр.197-200)
4. Соколов, А. В. Информационно-поисковые системы / А. В. Соколов. – М., 1981. – стр. 8-13, 40-60, 70-77.
5. Черный, А. И. Введение в теорию ИП / А. И. Черный. – М., 1975. стр. 25-100.
6. ГОСТ 7.25-2001 СИБИД. Тезаурус информационно-поисковый одноязычный. Состав, структура и основные требования к построению. http://www.complexdoc.ru/

Лекция 6

План:

I. Лексикография как одно из важных направлений прикладной лингвистики. Традиционная и машинная (компьютерная) лексикография. Основные направления компьютерной лексикографии.

II. Словарь как объект лексикографии. Классификация и организация словарей.

III. Идеографические словари и тезаурусы как примеры словарей.

I. ***Лексикография*** (от греч. *lexis 'слово'* и *grafia 'писание, наука'*), будучи одним из важных направлений прикладной лингвистики, занимается теорией и практикой составления словарей.

В лексикографии выделяют два научно-практических направления: *традиционная* и *машинная (компьютерная) лексикография*.

Традиционная лексикография имеет глубокие исторические корни и в большей мере занимается теорией и практикой составления «традиционных» словарей. Самые ранние ***глоссы*** (от греческого *glossa 'язык, слово'*, словарные пометы о значениях незнакомых слов) известны с глубочайшей древности (например, шумерские глоссы XXV в. до н. э.). Поэтому, самыми первыми типами словарей многие ученые считают *глоссарии*, написанные от руки списки иностранных и необычных слов, с которыми приходилось сталкиваться в манускриптах на древних языках.

Машинная лексикография *(Computational Lexicography)* занимается автоматизацией подготовки словарей и решает задачи разработки электронных словарей. В отличие от традиционной, машинная (компьютерная) лексикография – относительно молодая наука, реализующая традиционные наработки в технических средах и создающая разнообразные электронные словари. К основным задачам компьютерной лексикографии относятся также *задачи разработки технологий* составления электронных словарей и *управления терминологией* (*Terminology Management*).

В настоящее время можно выделить *три основных направления* компьютерной лексикографии:

1) автоматическое получение из текста различных словарей (например, терминологических, частотных словарей, словарей конкордансов и др.). Примеры страниц частотных словарей см. в Приложении 7;
2) создание словарей, являющихся электронными версиями традиционных словарей (например, словарь Даля), или комплексных электронных лингвистических словарей для традиционных словарных работ, например, известный словарь LINGVO [http://lingvopro.abbyyonline.com/ru]. Большой выбор такого рода открытых словарей доступен через поисковые системы, например, http://slovari.yandex.ru/.
3) разработка теоретических и практических аспектов составления специальных компьютерных словарей, например, для информационного поиска, машинного перевода (например, поисковые тезаурусы, словари стоп-слов, словари основ и флексий для морфологического анализа/синтеза). Словарь стоп-слов для информационного поиска приведен в Приложении 8.

Таким образом, мы видим, что основным объектом интересов лексикографии является такой продукт как словарь и все задачи, связанные с разработкой разного типа словарей.

II. **Словарь** – определенным образом организованное собрание слов, обычно с приписанными им комментариями, в которых описываются особенности их структуры и/или функционирования. Помимо слов, объектами словарного описания могут выступать их компоненты (таковы, например, словари морфем), словосочетания различных типов, устойчивые группы – пословицы, поговорки, цитаты и т. п. В другом значении термин «словарь», или *лексикон*, обозначает всю совокупность слов некоторого языка (иначе говоря, его лексику) и

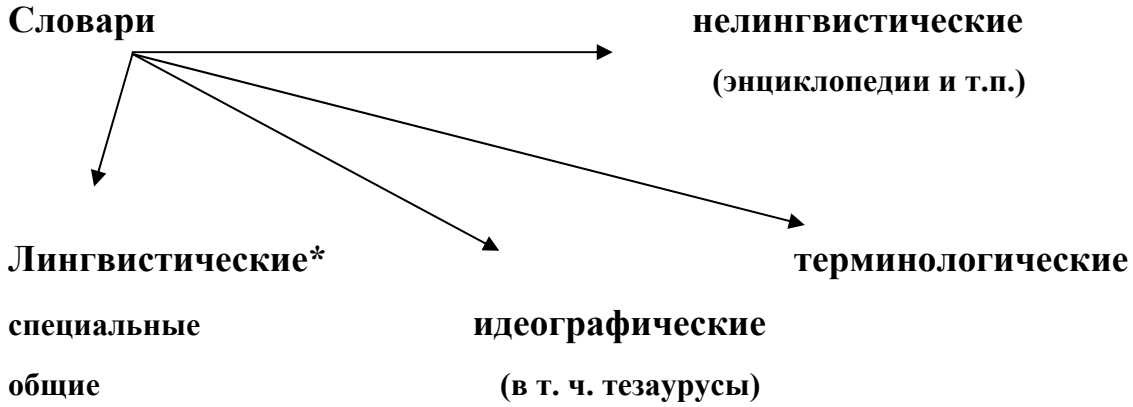
противопоставляется термину «грамматика», обозначающему совокупность правил построения из слов более сложных языковых выражений [1].

Таким образом, под ***словарем*** понимают:

- полный словарный состав языка;
- упорядоченное для решения практических задач множество лексических единиц;
- справочную книгу слов, расположенных в определенном порядке, дающую кому-либо информацию о том или ином слове.

Основная задача словаря – это представление либо описание лексики языка и ее особенностей для решения конкретных задач. Это сложнейшая проблема, так как лексика языка имеет тенденцию увеличиваться и качественно изменяться. Вот как метафорично говорит В. Селегей в статье «Электронные словари и компьютерная лексикография»: «*Многие словари, основной корпус статей которых сформировался в языковой атмосфере середины века, представляют собой лексикографические музеи (а то и терминологические кладбища, если говорить о специализированных словарях)»* [2].

Классификация словарей дана в Приложении 5, схематично можно представить ее следующим образом.



* *Примечание:* общие лингвистические словари (толковые, переводные, орфографические, частотные и др.), специальные (синонимов, антонимов, словарь пословиц и поговорок, сленга, словари иностранных слов и др.).

Основной структурной единицей словаря (как книги) является *словарная статья*, организация и моделирование которой является зачастую сложнейшей прикладной проблемой и задачей лексикографии.

Словник словаря – это перечень терминов словаря без их толкований.

Важным вопросом при составлении словарей также является порядок расположения словарных статей, чаще всего это алфавитный порядок или предметный (тематический), при котором слова группируются по темам или графически (например, тематический визуальный словарь).

Электронные словари часто представляют из себя сложный комплекс компьютерных программ – *лингвистических платформ*. Примером тому служит ABBYY Lingvo Content – профессиональное приложение, которое позволяет легко создавать новые словари и глоссарии, редактировать их и публиковать в бумажном и электронном виде на сайтах, на корпоративных порталах, а также в виде приложений для PC, Mac, смартфонов и мобильных устройств (http://www.abbyy.ru/lingvo_content).

Отметим некоторые особенности автоматических словарей:

- кроме словарной базы данных (перечень слов по алфавиту) для автоматической словарной работы необходимы специальные алгоритмы и программы, например, морфологической нормализации, или лемматизации. *Лемматизация* – это приведение разных форм слова к его канонической (исходной) форме (одна из задач компьютерной морфологии);
- в машинных словарях присутствуют не только перечни отдельных слов, но и до 50 % словосочетаний (особенно в терминологических словарях, которые очень важны для перевода специальных и технических текстов);
- электронные словари получили в настоящее время большое распространение в силу их доступности для широкого круга пользователей через различные мобильные и персональные устройства.

Примечание: следует отличать автоматические словари и системы машинного перевода (как ни странно, их часто путают обыватели). Последние лишь включают первые, являясь значительно сложнее.

III. Мы уже упоминали *словарь-тезаурус* как тип семантического словаря в предыдущей лекции. Однако тезаурусы имеют более широкие приложения, чем мы указали в лекции, посвященной информационному поиску.

Класс *идеографических словарей* (предметные, тематические), к которым часто относят и тезаурусы, – это особого рода словари, организованные, во-первых, по тематическому принципу, и, во-вторых, по принципу иерархии отношений и «от смысла к слову», т. е. идеографические словари ориентированы на семантику языка, и каждый такой словарь – это некоторая семантическая модель лексики, построенная на иерархических отношениях типа «род – вид», «часть – целое», «синонимы» и т. п.

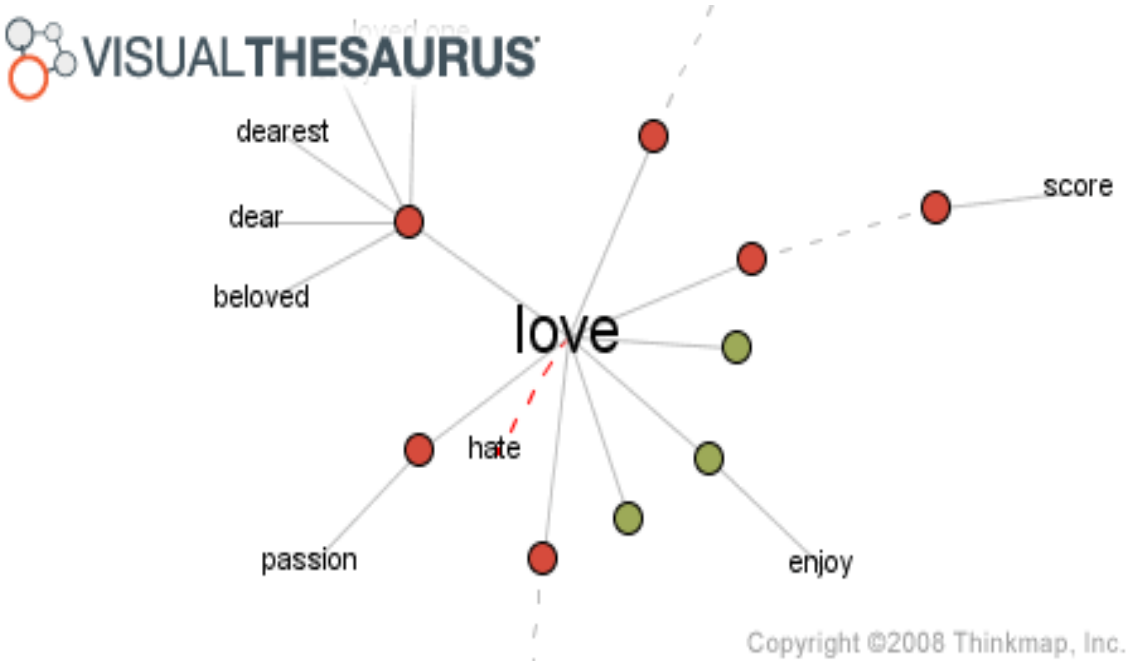
Вот что говорится в известном издании про идеографические словари «*Альтернативой алфавитному расположению слов является размещение их по смысловой близости. Словари, в которых лексика располагается на основании этого критерия, получили название* ***идеографических*** *(от греч. idea – понятие, идея, образ и grapho – пишу)*» [3].

Тезаурус – это также идеографический словарь, но имеющий четкую сложную иерархию отношений. В словарных статьях тезауруса отражены существенные для данного термина связи с другими понятиями, иными словами – это маленькая энциклопедия.

Самый известный классический словарь-тезаурус – это тезаурус П. Роже «*Thesaurus of English words and phrases»*, опубликованный в 1852 г., неоднократно переиздававшийся как классический лексикографический труд и представленный в сети Интернет [4, 5]. П. Роже систематизировал лексику английского языка по категориальным группам, каждая из которых представлена именем понятия, которых сначала насчитывалось 1000. Слова расположены в алфавитном порядке, далее идут синонимы слов по частям речи (существительные, глаголы, прилагательные, наречия), антонимы и затем списки родственных слов. Многими отмечается, что ценность этого тезауруса в

его естественности, в том, что это описание общей понятийной лексики языка, а также в том, что его можно привлекать к использованию в компьютерных системах как семантическое средство.

Направление, связанное с разработкой семантических словарей, активно развивается. Существуют глобальные проекты семантических словарных баз типа WORDNET [9], а также такие любопытные версии словарей как визуальные тезаурусы, наглядно представляющие визуальные карты выбранных слов, например [10]:



Семинар

План семинара

I. *Тезаурус П. Роже «Thesaurus of English words and phrases».*
II. *Терминография и терминологические словари. Управление терминологией как прикладная задача лингвистики.*

Домашнее задание

- *Сравнить модели словарной статьи академического издания – Словарь русского языка: В 4-х т. / Под ред. А. П. Евгеньевой и Энциклопедического издания «Слова о полку Игореве»: В 5 томах / Ред. кол.: Л. А. Дмитриев, Д. С. Лихачев, С. А. Семячко, О. В. Творогов (отв. ред.). – СПб.: Дмитрий Буланин, 1995. Т. 1. А–В. –*

1995. – 276 с. – Сделать вывод о различии лингвистического и энциклопедического словаря. Доступ – http://feb-web.ru/feb/feb/dict.htm

- *Через доступ к электронному каталогу библиотеки УлГТУ выбрать 10 видов словарей. Выбрать любые 3 разных словаря – проанализировать модель их словарной статьи.*
- *Выбрать определения понятия «термин» из книги – Шарафутдинова Н. С. Лингвокогнитивные основы научно-технической терминологии. – Ульяновск: УлГТУ, 2006. – 131 с.*

Список литературы

1. Энциклопедия Кругосвет – электронная версия – http://krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/SLOVAR.html (дата обращения: 30.08.2012)
2. В. Селегей. Электронные словари и компьютерная лексикография- http://www.lingvoda.ru/transforum/articles/selegey_a1.asp (дата обращения: 30.08.2012)
3. Морковкин, В. В. Идеографические словари / В. В. Морковкин. – М. : Изд-во МГУ, 1970. – http://rifmovnik.ru/ideog_book.htm (дата обращения: 30.08.2012)
4. Roget P. M. Thesaurus of English words and phrases. Электронная версия словаря – http://thesaurus.com/Roget-Alpha-Index.html (дата обращения: 30.08.2012)
5. Modern Language Association (MLA): "love." Roget's 21st Century Thesaurus, Third Edition. Philip Lief Group 2009. 14 Aug. 2012. <Thesaurus.com http://thesaurus.com/browse/love> (дата обращения: 30.08.2012)
6. Шарафутдинова, Н. С. Лингвокогнитивные основы научно-технической терминологии / Н. С. Шарафутдинова. – Ульяновск : УлГТУ, 2006. – 131 с.
7. Проект экспериментальных словарей – http://dict.ruslang.ru (дата обращения: 30.08.2012)

8. Словарь русского языка: В 4-х т. / РАН, Ин-т лингвистич. исследований / Под ред. *А. П. Евгеньевой*. – 4-е изд., стер. – М. : Рус. яз.; Полиграфресурсы, 1999. – http://feb-web.ru/feb/mas/mas-abc/default.asp (дата обращения: 30.08.2012)
9. WordNet® – http://wordnet.princeton.edu/ (дата обращения: 30.08.2012)
10. Visual thesaurus – http://www.visualthesaurus.com (дата обращения: 30.08.2012)

Лекция 7

План:
I. Машинный перевод. Виды МП.
II. Технологии МП. Задача машинного перевода как одна из важнейших задач прикладной лингвистики. Основные этапы системы МП, работающей на базе правил.

I. ***Перевод*** с одного языка на другой в общем случае состоит в изменении алфавита, лексики и синтаксиса языка с сохранением его семантики.

Перевод – это вид информационной деятельности, потребность в которой никогда не сокращается, а наоборот увеличивается. Исследования рынка переводов показали, что объемы этого вида деятельности постоянно увеличиваются, а в составе переводов преобладают специальные, научно-технические переводы – до половины общего объема рынка переводов, затем идут устный, учебный, синхронный, …, художественный.

Проблема моделирования перевода для приложения ее в компьютерной среде является центральной проблемой как прикладной лингвистики, так и искусственного интеллекта. Очевидно, что автоматизация перевода позволит повысить его эффективность, а также расширит границы межчеловеческой коммуникации.

Машинный перевод – это преобразование компьютером текста на одном естественном языке в эквивалентный по содержанию текст на другом естественном языке. Вот какое определение мы находим на сайте известной (и старейшей) системы машинного перевода SYSTRAN:

*«**Machine translation** (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Spanish). To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restored in the target language, i.e. the translation. While on the surface this seems straightforward, it is far more complex. Translation is not a mere word-for-word*

substitution. A translator must interpret and analyze all of the elements in the text and know how each word may influence another. This requires extensive expertise in grammar, syntax (sentence structure), semantics (meanings), etc., in the source and target languages, as well as familiarity with each local region» [1].

Системы МП – это целый комплекс специальных сложнейших программ и алгоритмов плюс специальные автоматические словари и формализованные грамматики для каждой языковой пары (входного и выходного языков).

История этого прикладного направления довольно хорошо задокументирована во многих источниках, например, о ней можно прочитать и кратко [2, 3], и целые тома [4, 5]. Историю МП отсчитывают с 1946 г., с возникновения такой науки, как кибернетика (наука об управлении). В 1954 г. был проведен так называемый Джоржтаунский эксперимент: переводился текст (250 слов) с русского на английский язык. Первая промышленная система МП SYSTRAN начинает функционировать в США в 1970 г. Первые опыты разработки коммерческих систем по МП показали огромные трудности при моделировании семантики языковых единиц и разрешении лингвистических неоднозначностей разного рода. Проблемы, связанные с моделированием всей структуры языка, полностью не решены до сих пор, поэтому были предложены другие подходы к организации машинного перевода, работающие на базе информации из больших объемов текстов, например, так появились технологии CAT – Computer-Aided-Translation, а также статистический перевод на базе поисковых систем, например, сервис Google Translate, работающий по собственным алгоритмам с 2007 г.

Виды современного МП

Существует несколько подходов к классификации систем машинного перевода, исходя из разных критериев.

Например, их часто делят в зависимости от степени участия человека в процессе перевода. Согласно этому критерию все системы делятся на **три типа**:

- **FAMT** (Fully-Automated Machine Translation) – полностью автоматический машинный перевод;
- **HAMT** (Human-Aided Machine Translation) – машинный перевод с участием человека;
- **MAHT** (Machine- Aided Human Translation) – перевод, осуществляемый человеком с привлечением вспомогательных программных и лингвистических средств.

Вот как различают эти термины авторы известного обзора «Survey of the State of the Art in Human Language Technology» конца XX века: *«…we would like to take the term machine assisted translation (MAT) as covering all techniques for automating the translation activity. The term* ***human-aided machine translation*** *(HAMT) should be reserved for the techniques which rely on a real automation of the translating function, with some human intervention in pre-edition, post-edition or interaction. The term* ***machine-aided human translation*** *(MAHT) concerns machine aids for translators or revisors…»* [6].

Мы также указывали на похожую классификацию в своем предыдущем издании этого пособия:

1) *информативный МП* – грубый пословный перевод без участия человека, достаточный для поверхностного ознакомления с содержанием текста (*автоматический*). Активно эксплуатируется и сейчас при поиске информации в поисковых системах;

2) *профессиональный МП* – более качественный перевод на базе коммерческих систем МП с последующим редактированием человеком (*автоматизированный*);

3) *интерактивный (персональный) МП* – считается переводом в специальных системах поддержки, например, в таких как CAT, эксплуатирующих базы перевода (TRANSLATION MEMORY), проходит в режиме диалога человека с компьютерной системой. Активно используется коммерческими компаниями.

Благодаря постоянному развитию технологий, накоплению лингвистических ресурсов и лингвистической статистики большинство современных систем МП дает неплохое качество перевода. Качество МП во многом зависит от возможностей настройки, имеющихся ресурсов (например, наличия готовых переводческих баз), от типа текстов (очевидно, что художественные тексты не имеет смысла переводить с помощью систем МП, а тексты инструкций к устройствам переводятся с хорошим качеством даже в автоматическом режиме).

Выше мы указали на одну из классификаций видов машинного перевода. Сейчас мы рассмотрим то, что часто называют *технологиями* или *подходами к осуществлению машинного перевода*. Здесь просматриваются существенные отличия и связано это, на наш взгляд, с тем, что разработчики систем машинного перевода, когда уровень и возможности современных аппаратных средств выросли, пошли по довольно эффективному пути минимизации творческих усилий, что проявилось в создании различных систем памяти переводов и лингвостатистических технологий.

II. Технологии машинного перевода

Как правило, выделяют следующие технологии, на базе которых работают современные системы машинного перевода:

- ***Example-based Machine Translation,*** или *Машинный перевод на базе готовых примеров переводов [7]*.

Эта концепция МП была предложена японским исследователем М. Нагао в 80-х годах XX века. Идея заключалась в следующем – *при накоплении достаточно большой коллекции ранее переведенных фраз для конкретных типов текстов и в узких областях велика вероятность того, что большая часть последующих текстов будет аналогична уже переведенным вручную*. Ярким представителем этого подхода являются инструменты CAT (*Computer-Aided-Translation tools*), реализующие технологии автоматизированной **поддержки перевода**.

К системам CAT относят те системы, которые обеспечивают работу на базе *«памяти переводов»* (*Translation Memory*), поэтому их часто называют *TM-инструменты* [8]. Функция обращения к базам «памяти перевода» дает переводчику и компании несомненные преимущества:

- Одинаковую лингвистическую единицу либо предложение, встречающиеся в разных местах, не нужно переводить дважды. Если подобное соответствие переведено, то необходимо лишь откорректировать предыдущий перевод.
- Если перевод текста выполняет группа переводчиков, то законченный текст получается более терминологически и стилистически однородным.

База данных TM состоит из пар параллельно сопоставленных *сегментов* (чаще всего это предложение), базы формируются программно по ходу перевода (изначально CAT-система пустая) и хранятся отдельно. CAT-программа при переводе производит обращение к базе, и если новый сегмент имеет точное или частичное соответствие в имеющейся накопленной базе переводов, то предлагается вариант перевода, если соответствия нет, то перевод идет вручную и вариант записывается в базу как новый сегмент.

Многими пользователями отмечается, что при использовании CAT-систем значительно повышается эффективность работы переводчика, работающего с типовыми специальными текстами, для которых уже накоплена богатая практика переводческих примеров, т. е. базы переводов.

Самыми популярными в России являются пакеты *SDL Trados, Deja Vu*. Кроме них существует большое количество как платных, так и бесплатных версий программ автоматизированного перевода, функционально программы похожи. Выходят демонстрационные бесплатные версии программ. Рынок CAT-программ постоянно развивается, параллельно накапливается множество специальных баз переводов.

- ***Statistical Machine Translation,*** или Статистический МП.

Эта технология машинного перевода также эксплуатирует идею вышеописанного EBMT-подхода, но имеет более строгую математическую базу, учитывая статистику тех лингвистических закономерностей, которые получены на базе структурного анализа текстов и анализа доступных параллельных корпусов текстов [7]. *Корпус параллельных текстов* – это тексты, содержащие предложения на одном языке и соответствующие им предложения на втором (например, двуязычные сайты). Статистический МП использует свойство «самообучения» языку (*machine learning*), т. е. чем больше накоплено параллельных текстов и чем точнее они соответствуют друг другу, тем лучше результат статистического машинного перевода.

Считается, что интерес к использованию лингвостатистических методов в МП возник еще в конце 40-х годов XX века параллельно с развитием теории информации К. Шеннона. В 60-х годах XX века исследователи успешно применяли лингвостатистику к решению задач дешифровки древних текстов. Но активно это направление стало развиваться только в начале 90-х годов XX века во время технологического и информационного бума, накоплением данных в корпусной лингвистике, машинном обучении и информационном поиске. Поэтому глобальные поисковые системы стали внедрять сервис статистического перевода в свои интерфейсы. Например, поисковик Google перешел на эту технологию в 2007 году и предлагает сервис Google Translate. В начале 2011 года Яндекс, как ранее Google, внедрил собственную подобную систему машинного перевода.

О принципах статистической технологии очень удачно и доходчиво изложено на сайте Яндекса. Приведем лишь выбранные отрывки в учебных целях (т. к. часто тексты полезных статей имеют тенденцию пропадать со временем на просторах Интернета):

«Такой перевод основывается не на правилах языка (системе эти правила даже не известны), а на статистике. Чтобы выучить язык, система сравнивает сотни тысяч параллельных текстов ... Это могут быть, например, большие тексты с разноязычных версий сайтов организаций…

В системе машинного перевода Яндекса три основные части: **модель перевода, модель языка и декодер.**

Модель перевода — это таблица, в которой для всех известных системе слов и фраз на одном языке перечислены все возможные их переводы на другой язык и указана вероятность этих переводов (для каждой пары языков есть своя таблица). Модель перевода создается в три этапа: сначала подбираются параллельные документы, потом в них – пары предложений, а затем уже пары слов или словосочетаний. … Система сравнивает не только отдельные слова, но и словосочетания из двух, трёх, четырёх или пяти слов, идущих подряд. В переводчике Яндекса модель перевода для каждой пары языков содержит более миллиарда пар слов и словосочетаний.

Другая составляющая системы машинного перевода — ***модель языка***. Для её создания система изучает сотни тысяч различных текстов на нужном языке и составляет список всех употребленных в них слов и словосочетаний с указанием частоты их использования. Это знание системы о языке, на который нужно перевести текст.

Непосредственно переводом занимается ***декодер***. Для каждого предложения исходного текста он подбирает все варианты перевода, сочетая между собой фразы из модели перевода, и сортирует их по убыванию вероятности. Например, пользователь захотел перевести фразу «to be or not to be». Допустим, из всех вариантов в модели перевода максимальная вероятность получилась у сочетания «быть или не бывает», сочетание «быть или не быть» оказалось с небольшим отрывом на втором месте и так далее. Все получившиеся варианты сочетаний декодер оценивает с помощью модели языка. В данном примере модель языка подскажет декодеру, что «быть или не быть» употребляется чаще, чем «быть или не бывает». В итоге декодер выбирает предложение с наилучшим сочетанием вероятности (с точки зрения модели перевода) и частоты употребления (с точки зрения модели языка)». *Подробнее:* http://company.yandex.ru/technologies/translation/

- Rule-based Machine Translation, или *Машинный перевод на базе лингвистических правил,* – это классическая технология МП, резко отличающаяся от предыдущих, с разработки которой и началась история МП и компьютерной лингвистики.

Глобальной задачей такой технологии машинного перевода является формальное моделирование естественных языков как объектов, реализованных в определенной программной и аппаратной компьютерной среде, а также моделирование различных процедур анализа, сопоставления (трансфера) и синтеза лингвистических единиц и текстов для этих естественных языков. Т. е. такая задача требует реализации процедур:

1) графематического (символьного) анализа;
2) морфологического и лексического анализа и синтеза;
3) синтаксического анализа и синтеза;

4) семантических преобразований.

При МП обрабатываются все уровни языковой структуры, переходя от одной подзадачи к другой. Для моделирования МП необходимо моделирование его подзадач, разработка алгоритмов и компьютерных программ для работы этих алгоритмов. Если внимательно присмотреться к этим задачам МП, то можно увидеть, что они представляют отдельные и сложные проблемы и направления прикладной лингвистики.

За историю машинного перевода как научного направления выделились три подхода к моделированию систем такого рода [8, 9]:

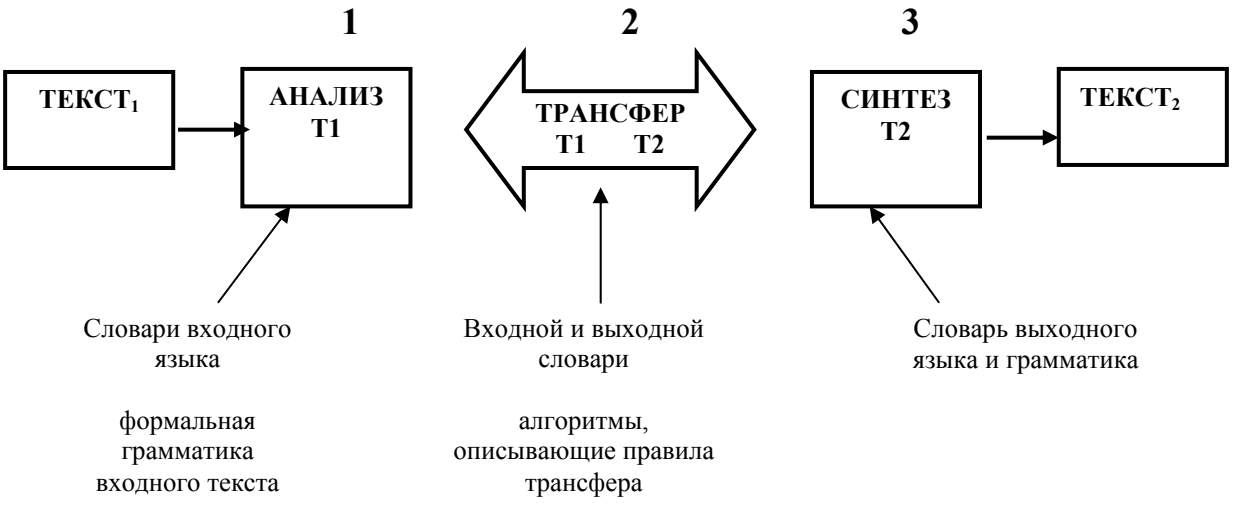
а) *«прямые»* системы МП давали пословный перевод (слово в слово);

б) системы «*интерлингвы»*, или *«текст–смысл–текст»,* целью было разработать смысловой язык-посредник;

Первый подход ограничен по своим возможностям, а для моделей-интерлингвы пока недостаточно научной базы и успешных разработок.

в) *«текст–трансфер-текст».* Перевод происходит на уровне переводных соответствий языковой пары. Единицей перевода выступает переводное соответствие, тесно связанное с лексической единицей (словосочетание, слово). В этой модели есть промежуточный этап – *трансфер*, на котором происходит установление переводных соответствий (согласование языковых единиц).

Классически МП разбивается на ***этапы*** анализа входного текста, трансфера и синтеза выходного текста [Марчук, с.106].



Преобразование Т1 должно начинаться с предварительной подготовки – его анализа. Обычно различают следующие виды лингвистического анализа: *морфологический, лексический, синтаксический, семантический.* Целью этапа анализа является получение всей лингвистической информации о выделенных единицах данного текста, например, получение всей морфологической информации о словоформе для последующего корректного перевода, и построение внутреннего представления входного предложения.

На этапе трансфера уже начинается переводческая работа, происходит сопоставление единиц по словарям, выполняются иные преобразования с помощью специальных формальных правил трансфера.

Цель этапа синтеза – на основе полученной в результате анализа информации построить (синтезировать) правильное предложение на выходном языке.

Эта модель построения *систем перевода на базе правил* наиболее удачная в коммерческом плане и в настоящее время преобладает над прямыми и интерлингвальными типами МП (на ней работают известные промышленные системы PROMT и SYSTRAN).

Мы кратко описали лишь базовые понятия и технологии современного машинного перевода, но это направление постоянно развивается, появляются новые способы автоматизации и повышения эффективности переводческой деятельности, различные гибридные технологии. Например, появляются технологии типа *Traduwiki (*http://traduwiki.org) для совместной переводческой деятельности, т. н. «облачные технологии» (*cloud-based translation)*, а люди продолжают придерживаться любимого ими, уже отработанного веками принципа «principle of least effort».

Семинар

План семинара

Практикум – Знакомство с компьютерными лингвистическими программами. Компьютерные переводчики - перевод небольшого текста примера, анализ ошибок машинного перевода.

Домашнее задание

- *Конспектирование работы - Harold Somers. Review Article: Example-based Machine Translation/ http://kitt.cl.uzh.ch/clab/satzaehnlichkeit/tutorial/Unterlagen/Somers 1999.pdf*
- *Протестировать машинный перевод систем ПРОМТ и Google Translate на русский язык 5 различных значений любого английского слова. Примеры контекста подобрать свои. Сделать вывод.*
- *Сделать компьютерную презентацию на 10 страниц (тема по выбору студента) по истории машинного перевода, какого-либо этапа развития МП, технологии МП, платформе МП.*

Список литературы

1. SYSTRAN – сайт системы http://www.systran.co.uk/systran/corporate-profile/translation-technology/what-is-machine-translation
2. Краткая история машинного перевода. Журнал «Русский репортер» – http://rusrep.ru/2010/24/istoriya_perevoda/
3. http://en.wikipedia.org/wiki/History_of_machine_translation
4. Yorick Wilks . Machine Translation: Its Scope and Limits. – Springer, 2008 – 254 p. - http://books.google.com/
5. William John Hutchins. Machine translation: past, present, future. - Ellis Horwood, 1986. – 382p.
6. Survey of the State of the Art in Human Language Technology – http://www.lt-world.org/hlt-survey/master.pdf (п. 8.3.1.)
7. Harold Somers. Review Article: Example-based Machine Translation/ http://kitt.cl.uzh.ch/clab/satzaehnlichkeit/tutorial/Unterlagen/Somers1999.pdf
8. Семенов, А. Л. Современные информационные технологии и перевод: учеб. пособие / А. Л. Семенов. – М. : Академия, 2008. – 224 с.
9. Марчук, Ю. Н. Методы моделирования перевода / Ю. Н. Марчук. – М., 1985. – С. 7-17, 43-45, 91-110, 135-137.

Лекция 8

План:

I. Текст и гипертекст.

Задачи автоматической обработки текста.

II. Современная лингводидактика.

Кратко рассмотрим другие прикладные задачи лингвистики, которые появились либо получили новое развитие с возникновением компьютерной техники.

I. С появлением компьютерных сетей и новых информационных технологий несколько трансформировались традиционные лингвистические понятия, в частности, понятие текста.

Под ***текстом*** (лат. *textus – связанность, материал, сплетение*), письменным или устным, принято понимать логически связанную последовательность лингвистических знаков. Основные характеристики текста: связанность; осмысленность; цельность (текст должен быть закончен по смыслу). Кроме того, классически текст имеет линейную структуру.

С середины 80-х годов XX века, когда быстрыми темпами стали развиваться компьютерные и телекоммуникационные сети (WWW, Internet) и вместе с тем глобальная коммуникация, в научный обиход вошло новое понятие *гипертекста* (*Hypertext*) [1]. Оказалось, что для решения многих информационных задач нелинейный и мультимедийный характер текста может быть эффективнее. ***Гипертекст*** – это технология организации информации и особым образом структурированный текст, разбитый на отдельные блоки, имеющий нелинейное представление, для эффективной презентации информации в компьютерных средах. Для стандарта представления информации в сетях стали разрабатываться специализированные языки. Например, HTML – *HyperText Markup Language* – язык гипертекстовой разметки документов.

При работе с текстом можно решать множество задач от общефилологических до статистических. Условно направление обработки

текста с помощью доступных компьютерных и информационных технологий можно назвать ***автоматической обработкой текста***, т. е. это любое преобразование текста на естественном или искусственном языках с помощью компьютера. Автоматическая обработка текстов осуществляется, как мы заметили раньше, в технологиях информационного поиска (при индексировании текстов), машинного перевода, словарной работе. В этом разделе мы рассмотрим некоторые неупомянутые ранее приложения и задачи.

- ***Автоматизированные операции по печати, редактированию и верстке текста***

Первоначально пользователю были доступны примитивные операции по обработке текста в простых компьютерных программах-редакторах (*Lexicon*), впоследствии требования к редактированию и представлению документов возрастали, что привело к созданию усовершенствованных систем типа WordProcessors (Microsoft Word и др.). Сейчас это развитые программы с функциями обработки таблиц, графики, *проверки орфографии и стилистики*. Кроме того, есть специальные издательские системы для профессиональной верстки документов, газет, рекламной продукции, книг (например, PageMaker или продвинутые пакеты типа Adobe® Digital Publishing Suite).

- ***Распознавание текста***

Распознавание текста в этом смысле привязано к задаче распознавания символов (*Character Recognition*) и реализуется эффективнее для печатного текста, чем для рукописного. С работой лингвистического распознавателя пользователь сталкивается при сканировании текста, например, через платформу ABBYY FineReader. В настоящее время задача распознавания печатного текста практически решена в силу ограниченного набора печатных гарнитур. Для рукописного текста (технологию называют ICR, т. е. *Intelligent Character Recognition*) распознавание графем гораздо сложнее, так как нужны более совершенные алгоритмы распознавания образов.

- ***Автоматическое реферирование и аннотирование текстов***

Автоматическое реферирование (Automatic Text Summarization) – это составление коротких изложений материалов, аннотаций или дайджестов, т. е. извлечение наиболее важных сведений из одного или нескольких текстовых документов и генерация краткого содержания анализируемого объекта. Аннотирование, в отличие от реферирования, предполагает еще большее сжатие содержания исходного текста. Существует много способов решения этой задачи прикладной лингвистики. Как правило, автоматическое реферирование основано на выборке ключевых фрагментов документов – выделении наиболее информативных фраз и элементов – и их сборке. При автореферировании активно используется статистическая информация по информативности разных элементов текста по частоте их появления в нем, информативность элемента текста также зависит от его позиции в документе или текстовых маркеров (например, термины выделены курсивом или жирно), которые характеризуют их содержательную значимость. Существуют как коммерческие, так и открытые, например, Open Text Summarizer [2], отдельные продукты для реферирования, а также встроенные функции автореферата (например, в MS Word).

- ***Корпусная лингвистика***

Корпусная лингвистика (*Corpus Linguistics*) – целое направление прикладной лингвистики, активно развивающееся с конца XX века, а в России – с начала XXI века, например, *Национальный корпус русского языка,* стал разрабатываться лишь в 2004 году [3]. В настоящее время электронных корпусов текстов для разных языков и приложений существует большое множество.

Задачи корпусной лингвистики связаны с разработкой технологий построения электронных лингвистических ресурсов особого типа – корпусов текстов (*Corpora*) – и решением задач разного рода на базе этих текстов. В основном, такие коллекции и массивы текстов отражают реальное функционирование того или иного языка, а их перенос в компьютерные среды

активизировал их практическое и широкое использование в прикладной лингвистике.

Корпусная лингвистика дает материал для различного рода исследований языка и его вариантов и определяет основной *метод анализа текстов и языка на базе корпусов*. Одной из важных особенностей метода анализа на базе корпусов является исследование не только чисто лингвистических явлений (грамматических или лексических функций слов, их связей с другими лексемами), но и таких явлений, как, например, частотности лексем или грамматических конструкций в тех или иных жанрах, диалектах.

Корпусный подход, или метод лингвистического исследования, основанный на корпусах текстов, ориентирован на прикладное изучение языка, его функционирования в реальных средах и текстах, что важно, например, для преподавания языка и для компьютерной лингводидактики, что затрагивается нами в последнем разделе курса.

II. ***Современная лингводидактика***

В первой лекции мы уже обращали с вами внимание на то, что англоязычный термин *«Applied Linguistics»* за рубежом часто трактуется несколько уже, чем мы привыкли в России, и под ним понимается направление, связанное с методами и технологиями преподавания языка, т. е. ***лингводидактика***. Этот взгляд на сферу интересов прикладной лингвистики существует и у нас в стране и поддерживается *Национальным обществом прикладной лингвистики*, НОПРИЛ, штаб которого находится в Московском государственном университете и руководит которым профессор С. Г. Тер-Минасова [5].

Одним из современных направлений прикладной лингвистики является возможность автоматизированного обучения иностранным языкам и поддержки его преподавания. За рубежом это направление, известное как ***Computer-Assisted Language Learning and Teaching*** (CALL/CALT), является перспективным в силу многих объективных причин и преподается как

специальная дисциплина прикладного языкознания на педагогических и лингвистических факультетах колледжей и университетов.

Применение компьютерных мультимедийных, дистанционных и открытых технологий в процессе обучения иностранным языкам положительно зарекомендовало себя в последние десятилетия [6]. Мультимедиа технологии позволяют моделировать среду, имитирующую лингвистическую и коммуникативную реальность, что очень важно для языкового обучения, а также активизировать основные методические принципы обучения иностранным языкам – развитие навыков: аудирования, говорения, чтения и письма. Неотъемлемой частью компьютерного образования с внедрением новых информационных и сетевых технологий стали такие электронные средства его поддержки, как машинные переводчики, словари, национальные корпусы текстов. Огромный потенциал для CALL/CALT несет доступ в Internet, в котором можно найти миллионы различных ресурсов для повышения эффективности иноязычного образования, а также различные технологии дистанционной коммуникации, например, Skype.

Семинар

План семинара

Практикум – Знакомство с компьютерными лингвистическими программами (лабораторный класс). Корпусы текстов – Британский национальный корпус (www.natcorp.ox.ac.uk) vs Национальный корпус русского языка (*http://www.ruscorpora.ru).*

Домашнее задание

- *Конспектирование работы – Ken Beatty. Teaching and Researching Computer-Assisted Language Learning (разделы - A brief history of CALL, Hypertext hypermedia and multimedia, Eight CALL applications) – доступ на books.google.com.*
- *Сделать простейший компьютерный тест из 25 вопросов по иностранному языку по любому разделу лексики или грамматики на базе HTML-шаблона, выданного преподавателем.*

- *На сайте http://training.abbyyonline.com/ выписать информацию о методике on-line обучения иностранному языку.*

Список литературы

1. Потапова, Р. К. Новые информационные технологии и лингвистика : учебное пособие / Р. К. Потапова – М. : МГЛУ, 2002. – 576 с.
2. Open Text Summarizer – http://libots.sourceforge.net/
3. Национальный корпус русского языка - http://www.ruscorpora.ru
4. Британский национальный корпус – www.natcorp.ox.ac.uk
5. Национальное общество прикладной лингвистики – http://www.nopril.ru/
6. Ken Beatty. Teaching and Researching Computer-Assisted Language Learning (разделы - A brief history of CALL, Hypertext hypermedia and multimedia, Eight CALL applications) – books.google.com

ЗАКЛЮЧЕНИЕ

Спецификой нашего учебного пособия является ориентация студентов на профиль практической деятельности будущих молодых специалистов в области теоретической и прикладной лингвистики. Курс лекций изначально дает установку на прикладной характер лингвистических исследований и сразу определяет круг основных задач этой предметной области. Рассмотренные в данном учебном издании задачи и сферы применения практических знаний специалистов не ограничены указанными, напротив, прикладная лингвистика является живой, интересной областью знаний, которая пополняется новыми лингвистическими приложениями и технологиями параллельно с развитием информационного общества.

Следует также отметить, что апробация авторского курса лекций прошла успешную проверку временем на базе кафедры «Прикладная лингвистика» Ульяновского государственного технического университета в течение последних 15 лет.

В заключение укажем, что по окончании вуза выпускники профиля теоретическая и прикладная лингвистика, как показывает опыт, могут применить полученные знания информационных технологий, иностранных языков и компьютерной лингвистики и работать в туристических и информационных агентствах, образовательных структурах, переводческих фирмах, в редакционных и информационных отделах организаций в качестве: лингвистов-переводчиков; разработчиков технической документации (технических писателей); специалистов информационных служб; специалистов по интернационализации и локализации изделий на международных рынках; SEO-оптимизаторов, разработчиков WEB-сайтов коммерческих фирм на иностранных языках.

На получение такого рода профессиональных компетенций и направлен наш профиль обучения.

СПИСОК ОБЯЗАТЕЛЬНОЙ ЛИТЕРАТУРЫ И ЭЛЕКТРОННЫХ РЕСУРСОВ

1. Баранов, А. Н. Введение в прикладную лингвистику : учебное пособие / А. Н. Баранов. – 2-е изд., испр. – М. : Едиториал УРСС, 2009. – 360 с.
2. Большой энциклопедический словарь. Языкознание / под ред. В. Н. Ярцевой – М. : Большая Российская энциклопедия, 1998. – 685 с. (см. соответствующие разделы и определения под терминами – социолингвистика, психолингвистика и т. д.).
3. Герд, А. С. Предмет и основные направления прикладной лингвистки. – http://project.phil.spbu.ru/lib/data/ru/heard/prikling.html (дата обращения: 08.08.2012)
4. Захаров, В. П. Информационные системы (документальный поиск) : учебное пособие / В. П. Захаров. – СПб. : Изд-во СПбГУ, 2002. – 188 с. – http://vp-zakharov.narod.ru/publications.htm (дата обращения: 11.08.2012)
5. Морковкин, В. В. Идеографические словари / В. В. Морковкин. – М. : Изд-во МГУ, 1970. – http://rifmovnik.ru/ideog_book.htm (дата обращения: 17.08.2012)
6. Селегей, В. Электронные словари и компьютерная лексикография - http://www.lingvoda.ru/transforum/articles/selegey_a1.asp (дата обращения: 17.08.2012)
7. Семенов, А. Л. Современные информационные технологии и перевод: учеб. пос. / А. Л. Семенов. – М. : Академия, 2008. – 224 с.
8. Потапова, Р. К. Новые информационные технологии и лингвистика : учеб. пос. / Р. К. Потапова. – М. : МГЛУ, 2002. – 576 с.
9. Шарафутдинова, Н. С. Лингвокогнитивные основы научно-технической терминологии / Н. С. Шарафутдинова. – Ульяновск : УлГТУ, 2006. – 131 с.
10. Шарафутдинова, Н. С. Теория и история лингвистической науки : учебник / Н. С. Шарафутдинова. – 3-е изд., испр. и доп. – Ульяновск : УлГТУ, 2012. – 139 с.
11. Этимологический словарь – Online Etymological Ductionary – http://www.etymonline.com/index.php?term=linguist (дата обращения: 23.08.2012)
12. Chomsky, Noam. Syntactic structures. Walter de Gruyter, 2002 – 117 с. – books.google.com (дата обращения: 30.08.2012)
13. Daniel Chandler. Semiotics for beginners (раздел Introduction) – http://www.aber.ac.uk/media/Documents/S4B/sem01.html (дата обращения: 30.08.2012)

14. Robert Marty. 76 Definitions of The Sign by C. S. Peirce – http://www.cspeirce.com/rsources/76defs/76defs.htm (дата обращения: 30.08.2012)
15. Roget P.M. Thesaurus of English words and phrases. Электронная версия словаря – http://thesaurus.com/Roget-Alpha-Index.html (дата обращения: 23.08.2012)
16. Survey of the State of the Art in Human Language Technology – http://www.lt-world.org/hlt-survey/master.pdf (дата обращения: 19.07.2012)
17. Harold Somers. Review Article: Example-based Machine Translation/ http://kitt.cl.uzh.ch/clab/satzaehnlichkeit/tutorial/Unterlagen/Somers1999.pdf (дата обращения: 28.07.2012)
18. Ken Beatty. Teaching and Researching Computer-Assisted Language Learning (разделы – A brief history of CALL, Hypertext hypermedia and multimedia, Eight CALL applications)- books.google.com (дата обращения: 28.07.2012)

Дополнительная рекомендуемая литература по курсу:

1. Герд, А. С. Прикладная лингвистика / А. С. Герд. – СПб. : Изд-во С.-Петерб. ун-та, 2005. – 268 с.
2. Звегинцев, В. А. Мысли о лингвистике / В. А. Звегинцев / серия «Из лингвистического наследия Звегинцева», 2008. – 336 с.
3. Зубов, А. В. Информационные технологии в лингвистике : учеб. пособие / А. В. Зубов. – М. : Академия, 2004. – 206 с. (Высшее профессиональное образование. Языкознание).
4. Искусственный интеллект : справочник: в 3 кн. / под ред. Э. В. Попова. – М. : Радио и связь, 1990. – Кн. 1: Системы общения и экспертные системы. – 464 с.
5. Искусственный интеллект : справочник: в 3 кн. / под ред. Д. А. Поспелова. – М. : Радио и связь, 1990. – Кн. 2: Модели и методы. – 303 с.
6. Кибрик, А. Е. Очерки по общим и прикладным вопросам языкознания (универсальное, типовое и специфичное в языке) / А. Е. Кибрик. – 3-е изд., стер. – М. : УРСС, 2002. – (Лингвистическое наследие 20 века). – 336 с.
7. Кандрашина, Е. Ю. Представление знаний о времени и пространстве в интеллектуальных системах / под ред. Д. А. Поспелова. – М. : Наука, 1989. – (Проблемы искусственного интеллекта; вып. 16). – 328 с. С. 8-34, 248-258.

8. Леонтьева, Н. Н. Автоматическое понимание текстов : учебное пособие / Н. Н. Леонтьева. – М. : Академия, 2006. – 304 с.
9. Марчук, Ю. Н. Методы моделирования перевода / Ю. Н. Марчук. – М., 1985. С. 7-17, 43-45, 91-110, 135-137.
10. Марчук, Ю. Н. Компьютерная лингвистика : учеб. пособие для студ. вузов, специализирующихся по направлению и спец. «Филология» / Ю. Н. Марчук. – М. : АСТ: Восток-Запад, 2007. – (Языкознание). – 317 с.
11. Марчук, Ю. Н. Основы терминографии : метод. пособие / Ю. Н. Марчук. – М. : ЦИИ МГУ, 1992. – 76 с.
12. Поспелов, Д. А. Моделирование рассуждений / Д. А. Поспелов. – М., 1989. – С. 124-151.
13. Потапова, Р. К. Речь: коммуникация, информация, кибернетика : учеб. пособие для вузов по спец. «Автоматизированные системы обработки информации и управления», «Лингвистика» / Р. К. Потапова. – М. : Радио и связь, 1997, 2001, 2010 – 528 с.
14. Прикладное языкознание : учебник / Л. В. Бондарко и др. – СПб.,1996. – 528 с.
15. Соколов, А. В. Информационно-поисковые системы / А. В. Соколов. – М., 1981. – С. 8-13, 40-60, 70-77.
16. Черный, А. И. Введение в теорию ИП / А. И. Черный. – М., 1975. – С. 25-100.

Дополнительные электронные ресурсы

(часть ресурсов также указана в тексте пособия):

17. www.dialog-21.ru/ (общая информация по направлениям КОЛИНТ)
18. Сайт кафедры ОТИПЛ МГУ(общая информация)
http://www.philol.msu.ru/~otipl/new/main/index.php
19. Открытая электронная энциклопедия «Википедия» (запросы по темам курса), например, http://en.wikipedia.org/wiki/Linguistics,
http://ru.wikipedia.org/wiki/Linguistics
20. Открытая электронная энциклопедия «Кругосвет» (запросы по темам курса) - http://www.krugosvet.ru
21. Свободная энциклопедия языков программирования - http://progopedia.ru/

22. Пирс Ч. Прагматизм. Научная метафизика.
- http://filosof.historic.ru/books/item/f00/s00/z0000699/st007.shtml
23. ГОСТ 7.25-2001 СИБИД. Тезаурус информационно-поисковый одноязычный. Состав, структура и основные требования к построению. http://www.complexdoc.ru/
24. ГОСТ Р7.24-2007 СИБИД. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению. – 2008 г. http://www.lib.tsu.ru/index_main.php?id=11
25. Проект экспериментальных словарей – http://dict.ruslang.ru
26. SYSTRAN – сайт системы http://www.systran.co.uk
27. Краткая история машинного перевода. Журнал «Русский репортер» – http://rusrep.ru/2010/24/istoriya_perevoda/
28. http://en.wikipedia.org/wiki/History_of_machine_translation
29. Yorick Wilks . Machine Translation: Its Scope and Limits. – Springer, 2008 – 254 p. – http://books.google.com/
30. William John Hutchins. Machine translation: past, present, future. – Ellis Horwood, 1986. – 382p.
31. Claire Bradin Siskin. CALL on the Web – http://edvista.com/claire/call.html
32. Open Text Summarizer – http://libots.sourceforge.net/
33. Национальный корпус русского языка – http://www.ruscorpora.ru
34. Британский национальный корпус – www.natcorp.ox.ac.uk
35. Национальное общество прикладной лингвистики – http://www.nopril.ru/
36. Фундаментальная электронная библиотека «Русская литература и фольклор» (ФЭБ) – http://feb-web.ru/

Дополнительные материалы по курсу также представлены в электронном виде на сервере лаборатории 304 в каталоге 1 курса и предоставляются по запросу студента.

ПРИЛОЖЕНИЕ 1

Рабочая программа по курсу

Содержание дисциплины формируется за счет тематических блоков, изучаемых последовательно в ходе лекционных и семинарских занятий, и представлено в таблицах. Академические часы и часы для самостоятельной работы студента могут корректироваться согласно учебному плану и изменению объема программы курса. Домашние задания указаны после каждого раздела курса, домашние задания могут меняться и конкретизироваться.

Теоретический курс

Таблица 1

Раздел, тема учебной дисциплины, содержание темы	**Номер лекции**	**Количество часов**	
		лекции	**СРС**
Тема 1. 1.1. Введение. Лингвистика как наука. Система лингвистических дисциплин и направлений. Различные подходы к определению термина «Прикладная лингвистика» (Applied Linguistics, Computational Linguistics). 1.2. Теоретическая и прикладная (практическая) лингвистика. Задачи и направления прикладной лингвистики. 1.3. Связь лингвистики с другими науками – естественными и гуманитарными.	1	2	2
Тема 2. 2.1. Семиотика. Взаимосвязь с лингвистикой. 2.2. Знак и классификации знаков. 2.3. Язык как знаковая система. 2.4. Информация. Виды и представление информации. Получение, хранение, обработка и передача вербальной информации. Информатика как наука. Взаимосвязь с лингвистикой. 2.5. Компьютерная лингвистика.	2	2	4
Тема 3. 3.1. Естественные и искусственные языки. Языки человеко-машинного общения и программирования как искусственные языки. 3.2. Формальные методы описания искусственных языков. Понятие формальной грамматики и языка. Понятие метаязыка. БНФ-нотации и синтаксические диаграммы. 3.3. Математическая лингвистика и теория формальных языков.	3	2	4

Тема 4. 4.1. Моделирование как основной метод в прикладной и компьютерной лингвистике. Понятия «модель» и «лингвистическая модель». 4.2. Модели знаний и семантики в прикладной лингвистике и искусственном интеллекте. зыки представления знаний как вариант искусственных языков. Их классификация.	4	2	14
Тема 5. 5.1. Информационный поиск. Виды поиска, характеристики информационного поиска. Лингвистика в задачах информационного поиска. 5.2. Информационно-поисковые языки как искусственные языки. Классификация информационно-поисковых языков.	5	2	4
Тема 6. 6.1. Лексикография как одно из важных направлений прикладной лингвистики. Словарь как объект лексикографии. Классификация и организация словарей. 6.2. Традиционная и машинная лексикография. Основные направления компьютерной лексикографии. 6.3. Некоторые особенности автоматических словарей. Идеографические словари и тезаурусы. 6.4. Терминография. Терминологические банки данны.	6	2	4
Тема 7. 7.1. Машинный перевод. Задача машинного перевода как одна из важнейших задач прикладной лингвистики. Хронология развития. 7.2. Виды МП. 7.3. Подходы к моделированию МП (прямой, через интерлингву, через трансфер). 7.4. Технологии МП. Основные этапы работы системы МП. Соотношение анализа и синтеза текста. Понятие трансфера.	7	2	4
Тема 8. 8.1. Автоматическая обработка текста. Текст и гипертекст. 8.2. Компьютерная лингводидактика. 8.3. Перспективные направления лингвистики (корпусная лингвистика, политическая лингвистика, когнитивная, медиалингвистика и др.).	8	2	14
Итого	**8**	**16**	10 экза- мен **60**

Продолжение ПРИЛОЖЕНИЯ 1

Практические (семинарские) занятия

Таблица 2

Номер занятия	Наименование темы и содержание занятия	Номер раздела, тема дисциплины	Формы контроля выполнения работы	Объем в часах	
				Аудиторных	СРС
1	**2**	**3**	**4**	**5**	**6**
1	Понятие о языке как средстве коммуникации. Речь и язык – предмет и объект языкознания. Структурные уровни языка. Языковые системы.	Тема 1	опрос	2	1
2	Семантический треугольник. Информация – понимание в разных науках.	Тема 2	Типовая работа 1	2	5
3	Линейные и ветвящиеся синтаксические диаграммы. Пример БНФ нотации. Примеры формальных грамматик и вывода на них.	Тема 3	Типовая работа 2	2	5
4	Эвристические модели знаний. Сетевые модели. Пример семантических сетей и вывода на них. Фрейм как шаблон типовой ситуации. Сценарии на примерах.	Тема 4	Типовая работа 3	2	5
5	Логические средства при представлении знаний. Логика предикатов. Отношения для представления лингвистических знаний и информации (на примере временной логики).	Тема 4	Типовая работа 4	2	5
6	Информационный поиск. Хранение информации в ИПС. Записи информационно - поисковых языков. Тезаурус → дескрипторная статья. Словари → определение типа словаря.	Тема 5, 6	Типовая работа 5	2	5
7	Знакомство с компьютерными лингвистическими..программами в дисплейном классе	Тема 6, 7	Практикум (в компьютерном классе)	2	8
8	Лингвистика и новые информационные технологии (демонстрация в компьютерном классе)	Тема 8	Презентация или практикум	2	8
Итого				**16**	**42**

Литература по курсу, ссылки на цитаты и ссылки на работы для конспектирования указаны в тексте настоящего пособия.

ПРИЛОЖЕНИЕ 2

Примеры теоретических и тестовых вопросов для контроля знаний по прикладной лингвистике

Раздел 1.

1.1. Лингвистика как наука. Система лингвистических дисциплин и направлений. Различные подходы к определению термина «Прикладная лингвистика» (Applied Linguistics, Computational Linguistics).

1.2. Теоретическая и прикладная (практическая) лингвистика. Задачи и направления прикладной лингвистики.

1.3. Связь лингвистики с другими науками – естественными и гуманитарными. Связь лингвистики с филологией. Новые информационные и компьютерные технологии в прикладной лингвистике.

- **Объектом изучения прикладной лингвистики является**…….

- **В зарубежной лингвистике термин «Applied Linguistics», «Angewandte Linguistik» часто связывается с:**
  – автоматической обработкой речи
  – преподаванием иностранных языков
  – автоматической обработкой информации
  – машинным переводом

- **Соответствие трех глобальных направлений языкознания:**

теоретическое	- конкретное описание языковых явлений.
описательное	- совершенствование языковой системы.
прикладное	– объяснение языковых систем и процессов.

Раздел 2.

2.1. Семиотика. Взаимосвязь с лингвистикой.

2.2. Знак и классификации знаков.

2.3. Язык как знаковая система.

2.4. Информация. Виды и представление информации. Получение, хранение, обработка и передача вербальной информации. Информатика как наука. Взаимосвязь с лингвистикой.

2.5. Компьютерная лингвистика.

- **Наука, изучающая знаковые системы** –
  – дешифровка
  – грамматика
  – семиотика
  – семантика

- **Основателем семиотики считается**
  – Ч.Моррис
  – Ч.Пирс
  – Ф. де Соссюр
  – Н. Хомский

- **Переводным синонимом термина «компьютерная лингвистика» является:**
  – computer linguistics
  – computational linguistics
  – corpus linguistics

Раздел 3.
3.1. Естественные и искусственные языки. Языки человеко-машинного общения и программирования как искусственные языки.
3.2. Формальные методы описания искусственных языков. Понятие формальной грамматики и языка. Понятие метаязыка. БНФ-нотации и синтаксические диаграммы.
3.3. Математическая лингвистика и теория формальных языков.

- **……** – это знаковые системы, создаваемые для использования в тех областях науки и техники, где применение естественного языка ограничено, менее эффективно или невозможно.

- **Исторически языки программирования возникли**
  – в 60-х гг. XX вв.
  – в 40-х годах XX вв.
  – в 80-х гг. XX вв.

- **Вопросами формальных грамматик и их теорией занималась** …
  – математическая лингвистика
  – структурная лингвистика
  – прикладная лингвистика
  – лингвостатистика

Продолжение ПРИЛОЖЕНИЯ 2

Раздел 4.

4.1. Моделирование как основной метод в прикладной и компьютерной лингвистике. Понятия «модель» и «лингвистическая модель».

4.2. Модели знаний и семантики в прикладной лингвистике и искусственном интеллекте.

4.3. Языки представления знаний как вариант искусственных языков. Их классификация.

- **«Искусственно создаваемое лингвистом реальное или мысленное устройство, воспроизводящее, имитирующее своим поведением (обычно в упрощенном виде) поведение оригинала в лингвистических целях», – это определение.....**

– языка моделирования
– лингвистической модели
– модели речевой деятельности
– метаязыка

- **В основе …….. лежит понятие структуры, образованной помеченными вершинами и дугами. Вершины представляют некоторые сущности, а дуги – отношения между связываемыми ими сущностями.**

– логической модели
– сетевой модели
– фреймовой модели
– сценария

- **Знания традиционно разделяются на два вида :**

– описательное и теоретическое
– процедурное и фактографическое
– декларативное и процедурное
– понятийное и декларативное

Раздел 5.

5.1. Информационный поиск. Виды поиска, характеристики информационношгго поиска.

5.2. Информационно-поисковые языки как искусственные языки. Классификация информационно-поисковых языков. Поисковый тезаурус.

- **В зависимости от объекта хранения и типа запроса различают два вида информационного поиска:**

– документальный и расширенный
– документальный и ключевой
– документальный и фактографический
– фактографический и расширенный

- **Документ, центральный предмет или тема которого в целом соответствует смысловому содержанию информационного запроса, называется ……**

– смысловым
– индексом
– ссылкой
– релевантным

- **Индексирование при автоматическом поиске информации – это…..**

– поиск и выделение ключевых слов
– поиск релевантных документов
– формальное описание содержание документа на ИПЯ
– составление запроса на поиск документа

Раздел 6.

6.1. Лексикография как одно из важных направлений прикладной лингвистики. Словарь как объект лексикографии. Классификация и организация словарей.
6.2. Традиционная и машинная лексикография. Основные направления компьютерной лексикографии.
6.3. Некоторые особенности автоматических словарей. Идеографические словари и тезаурусы.
6.4. Терминография. Терминологические банки данных.

- **Это важное направление прикладной лингвистики занимается теорией и практикой составления словарей……**

- **Тезаурус – это вид ……**

– частотного словаря
– идеографического словаря
– синтаксического словаря

- **….. – это приведение разных форм слова к его канонической (исходной) форме.**

– индексирование
– лемматизация
– стемминг
– нормализация

Раздел 7.

7.1. Машинный перевод. Задача машинного перевода как одна из важнейших задач прикладной лингвистики. Хронология развития.
7.2. Виды МП. Их практическое применение.

7.3. Подходы к моделированию МП (прямой, через интерлингву, через трансфер).
7.4. Технологии МП.
7.5. Основные этапы работы системы МП. Соотношение анализа и синтеза текста. Понятие трансфера.

- **……** – это преобразование компьютером текста на одном естественном языке в эквивалентный по содержанию текст на другом естественном языке.

- **Первое крупное событие в истории машинного перевода – Джоржтаунский эксперимент – произошло …**
  – в 1949 г.
  – в 1954 г.
  – в 1966 г.

- **Наиболее известной зарубежной системой машинного перевода, работающей по правилам, является:**
  – ProMT
  – FOBOS
  – SYSTRAN
  – Google

Раздел 8.
8.1. Автоматическая обработка текста. Текст и гипертекст.
8.2. Лингвостатистика в задачах обработки текста.
8.3. Компьютерная лингводидактика.
8.4. Перспективные направления лингвистики (корпусная лингвистика, политическая лингвистика, когнитивная, медиалингвистика и др.).

- «Особым образом структурированный текст, разбитый на отдельные блоки, имеющий нелинейное представление, для эффективного представления информации в компьютерных средах» – **это определение……..**

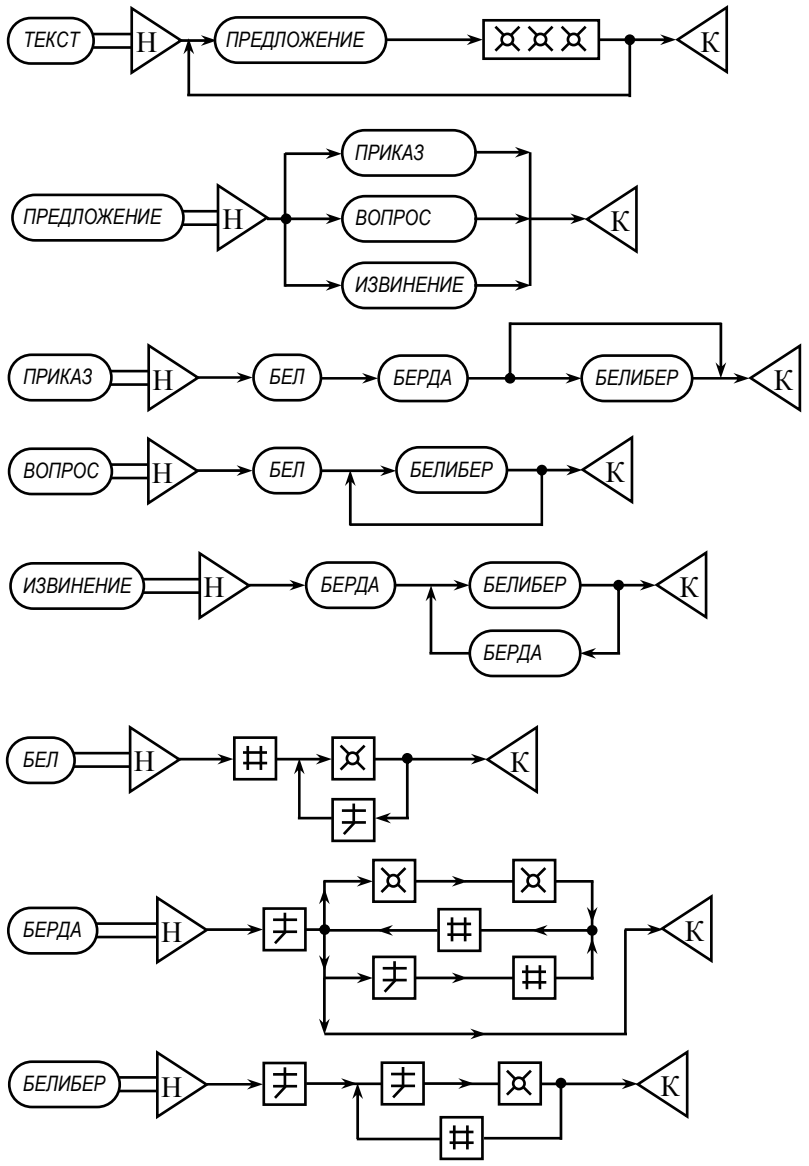
- **Научное направление, изучающее проблему моделирования языка и мышления в компьютерных средах (например, для робототехники), называется ……**
  – компьютерная лингвистика
  – психолингвистика
  – искусственный интеллект
  – когнитивная лингвистика

- **Понятием «Corpora» в корпусной лингвистике определяют……**

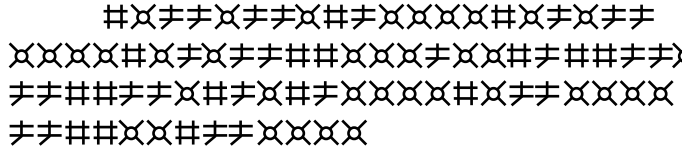
ПРИЛОЖЕНИЕ 3

Примеры задач к экзамену по курсу

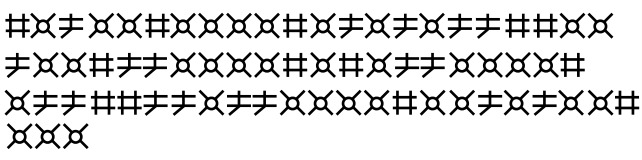
1. Алфавит белибердинского языка состоит из трех символов. На основе анализа многочисленных надписей ученым удалось установить, что синтаксис языка описывается диаграммой:



Смысл предложений языка не выяснен. На экспертизу в лабораторию поступили две новые надписи. Предполагается, что одна из них содержит подлинный текст, а другая является подделкой. Постарайтесь провести экспертизу и обнаружить подделку, если она есть.



а)



б)

Проведите синтаксический разбор подлинного текста и определите, сколько в нем приказов, сколько вопросов и сколько извинений.

2. Построить семантический треугольник для следующих лингвистических символов: а) перевод; б) информация; в) словарь. Дать представление этих понятий в виде словарной статьи тезауруса.

3. Составить БНФ-нотацию для издания справочника по импортным лекарственным средствам (справочник взять у экзаменатора).

4. Построить синтаксическую диаграмму для текста:

«Why do you cry, Billy?

Why do you cry?

Why, Billy?

Why, Billy?

Why, Billy?

Why?»

5. Дана некоторая формальная грамматика (пример у экзаменатора). Какой язык она порождает? Какая конструкция семантически неправильная? Как записать правила, чтобы избежать ошибки?

ПРИЛОЖЕНИЕ 4

Пример визуальной семиотики текстов

Рисунок. Визуальная семиотическая карта частотной лексики фишинговых спам-сообщений *

*** Источник** –

Соснина. Е.П. Бурлука К.В.. *Лингво-семиотические характеристики фишинговых сообщений:/* сборник материалов конференции: РОССИЙСКАЯ НАУЧНАЯ КОНФЕРЕНЦИЯ С МЕЖДУНАРОДНЫМ УЧАСТИЕМ«Системный анализ и семиотическое моделирование (SASM’2011), КАЗАНЬ 24-28 февраля 2011. – Казань, 2011.

ПРИЛОЖЕНИЕ 5

Примеры программ на языке программирования

Текст программы на языке программирования Паскаль

Исходное задание – *В заданном файле primer.txt сделать анализ начальных букв англоязычных слов, построив из символов * простейшую гистограмму (график частотности) найденных начальных согласных. Вывести на печать слова, которые начинаются с согласной буквы, которая чаще всего встречается.*

```
program primer;
uses crt;
type Letters = set of 'A'..'Z';
TFreqs = array ['A'..'Z'] of word;
const Vowels : Letters = ['A', 'E', 'I', 'O', 'U', 'Y'];
var f : Text;
freqs : TFreqs;
max_freq,pos : word;
s : string;
c_pos,freq_let : char;
begin
clrscr;
assign(f,'c:\primer.txt');
reset(f);
while not eof(f) do
   begin
readln(f,s);
for pos := 1 to length(s) do
if ((pos = 1) OR (s[pos-1] = ' ')) AND not (upcase(s[pos]) in Vowels) then
inc(freqs[upcase(s[pos])]);
   end;
for c_pos := 'A' to 'Z' do
if freqs[c_pos] <> 0 then
  begin
write('',c_pos,' ');
for pos := 1 to freqs[c_pos] do write('*');
writeln;
  end;
max_freq := freqs['A'];
for c_pos := 'B' to 'Z' do
if freqs[c_pos] > max_freq then
  begin
max_freq := freqs[c_pos];
freq_let := c_pos;
  end;
close(f);
reset(f);
writeln(#13#10,'Words with first letter ',freq_let,':',#13#10);
while not eof(f) do
   begin
readln(f,s);
for pos := 1 to length(s) do
if ((pos = 1) OR (s[pos-1] = ' ')) AND (upcase(s[pos]) = freq_let) then
     begin
write('');
```

```
while (s[pos] <> ' ') AND (pos <> length(s)) do
      begin
write(s[pos]);
inc(pos);
     end;
writeln("");
   end;
  end;
readln;
end.
```

Текст программы на языке программирования PHP

Исходное задание – *Создать скрипт регистрации посетителей on-line в таблице session**

```
<?php
  // Начинаем сессию
  session_start();
  // Получаем уникальный id сессии
  $id_session = session_id();
  // Устанавливаем соединение с базой данных  include "config.php";
  // Проверяем, присутствует ли такой id в базе данных
  $query = "SELECT * FROM session
          WHERE id_session = '$id_session'";
  $ses = mysql_query($query);
  if(!$ses) exit("<p>Ошибка в запросе к таблице сессий</p>");
  // Если сессия с таким номером уже существует,  значит пользователь online -
   // обновляем время его последнего посещения
  if(mysql_num_rows($ses)>0)
  {
    $query = "UPDATE session SET putdate = NOW(),
                            user = '$_SESSION[user]'
             WHERE id_session = '$id_session'";
    mysql_query($query);
  }
  // Иначе, если такого номера нет - посетитель только что
  // вошёл - помещаем в таблицу нового посетителя
  else
  {
    $query = "INSERT INTO session
            VALUES('$id_session', NOW(), '$_SESSION[user]')";
    if(!mysql_query($query))
    {
      echo $query."<br>";
      echo "<p>Ошибка при добавлении пользователя</p>";
      exit();
    }
  }
  // Будем считать, что пользователи, которые отсутствовали
  // в течении 20 минут - покинули ресурс - удаляем их  id_session из базы данных
  $query = "DELETE FROM session
          WHERE putdate < NOW() -  INTERVAL '20' MINUTE";
  mysql_query($query);
?>
```

* пример кода с сайта - http://www.softtime.ru/scripts/online.php

ПРИЛОЖЕНИЕ 6

«ТИП СЛОВАРЯ» –

основные фасеты для характеристики

Ф 1. Тип словаря по принципу определения значения единиц словника. В основу этого фасета положено традиционное противопоставление словарей: энциклопедии – обычные (филологические) словари. Но в связи с тем, что некоторые типы общих (филологических) словарей включают в описание слов элементы энциклопедизма, целесообразно (наряду с признаками: энциклопедический, филологический) ввести дополнительный признак – энциклопедично-филологический.

Строгое противопоставление словарей: энциклопедические – филологические (лингвистические) – обязательно далеко не во всех случаях. Отсутствие такого строгого деления можно видеть в терминологических словарях, где объяснение значения термина незначительно отличается от краткого описания соответствующего денотата.

Ф 2. Тип словаря по способу организации лексики: алфавитный, гнездовой, словопроизводный, тематический (идеографический, идеологический), алфавитно-тематический, частотный, обратный.

Ф 3. Тип словаря по объему словника: тезаурус-сокровищница, большой (полный, интегральный), малый (краткий, выборочный) словарь.

Ф 4. Тип словаря по отношению к типу языкового общения: общий (включающий лексику современного литературного языка, частично диалектную, устаревшую, разговорную и терминологическую), областной, словарь языка писателя, терминологический (технический, научно-технический), научно-специальный, словари для описания фразеологии, синонимии, антонимии.

Ф 5. Тип словаря по отношению к языковой норме: нормативный – ненормативный. Нормативный словарь служит руководством к правильному употреблению слов, к правильному образованию их форм, правильному

произношению и правильному написанию. Ненормативный словарь описывает то, как реально говорят, а не то, как надо говорить.

Ф 6. Тип словаря по способу представления лексического значения слова: одноязычный – многоязычный (толковый – переводной). Одноязычный (толковый) словарь описывает лексику способами данного языка, используя описательно-логический метод. Многоязычный (переводной) словарь (двуязычный, трехъязычный и др.), являясь не чем иным, как описанием лексики одного языка при помощи лексических средств другого языка, использует обычно синонимический способ перевода значения слова с одного языка на другие.

Ф 7. Тип словаря по отношению к синхронии и диахронии: исторический, неисторический, этимологический. Исторический словарь (в широком смысле) – это словарь, характеризующий жизнь слов на протяжении определенного периода истории развития языка. Такого словаря для русского языка не существует.

Многие лингвисты в разряд исторических словарей относят словари, которые не ставят целью описать историю слов, а описывают значения слов, зафиксированных в литературных памятниках определенной эпохи. Эти словари по принципу описания значения слов не отличаются от нормативных словарей современного языка, так как описывают лексику в синхронном срезе.

Ф 8. Тип словаря по назначению (цели): справочник, учебный, информационно-поисковый. Специфика учебных словарей заключается, прежде всего и главным образом, в их учебной направленности. Этот фактор определяет и состав лексики (отбор слов), и особенности ее описания. Основным назначением информационно-поисковых словарей (тезаурусов для определенной области науки) является составление с их помощью поисковых образов документов и запросов при вводе их в информационную систему.

Классификационная схема понятия «тип словаря» построена на основе восьми перечисленных выше оснований деления.

Тип словаря

Ф1 по принципу определения значения единиц словника:

филологический,

энциклопедический,

энциклопедично-филологический;

Ф2 по способу организации лексики словаря:

алфавитный,

гнездовой,

обратный,

словопроизводный,

тематический,

идеографический,

идеологический,

алфавитно-тематический,

частотный;

Ф3 по объему словника:

большой,

полный,

интегральный,

малый,

выборочный,

краткий,

средний,

тезаурус-сокровищница;

Ф4 по отношению к типу языкового общения:

общий,

словарь литературного языка,

областной,

диалектный,

словарь народных говоров,

словарь языка писателя,

терминологический,

технический,

научно-технический,

научно-специальный,

словарь для описания специальных средств языка,

антонимический,

синонимический,

фразеологический,

омонимический;

Ф5 по отношению к языковой норме:

ненормативный,

нормативный,

орфографический,

орфоэпический,

стилистический;

Ф6 по способу представления лексического значения слова:

многоязычный,

переводной,

двуязычный,

трехъязычный и т. д.,

одноязычный толковый;

Ф7 по отношению к синхронии и диахронии:

исторический,

неисторический,

этимологический;

Ф8 по назначению (цели):

справочник,

учебный,

информационно-поисковый.

Конкретный словарь будет описываться набором **признаков**, взятых из каждого фасета. Например, «Словарь иностранных слов» под ред. Н. В. Лехина (М., 1994) – энциклопедично-филологический, алфавитный, краткий, терминологический, нормативный, одноязычный (толковый), неисторический, справочник.

ПРИЛОЖЕНИЕ 7

Примеры частотного словаря

Ниже исключительно в ознакомительных целях с типом словаря приведем примеры из замечательного частотного словаря, составленного на базе Национального корпуса русского языка, *О. Н. Ляшевская, С. А. Шаров, Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. Электронная версия издания: http://dict.ruslang.ru/freq.php*

Частотный список имен существительных

	Лемма	Частота (ipm)	Ранг
1	год	3727.5	28
2	человек	2723.0	39
3	время	2015.7	52
4	дело	1412.1	65
5	жизнь	1389.8	66
6	день	1258.4	71
7	рука	1200.6	74
8	раз	1131.8	78
9	работа	1058.3	87
10	слово	967.9	94
11	место	926.6	98
12	лицо	878.0	104
13	друг	874.2	106
14	глаз	857.6	110
15	вопрос	805.8	114
16	дом	792.6	118
17	сторона	768.3	121
18	страна	725.7	125
19	мир	714.7	128
20	случай	709.7	131

Частотный список глаголов

	Лемма	Частота (ipm)	Ранг
1	быть	12160.7	6
2	мочь	2912.3	37
3	сказать	2396.6	42
4	говорить	1755.0	58
5	знать	1713.8	61
6	стать	1621.8	62
7	хотеть	991.3	92
8	идти	957.1	95
9	иметь	906.7	99
10	видеть	818.2	113
11	думать	755.5	123
12	сделать	743.5	124
13	жить	725.5	126
14	делать	701.1	134
15	смотреть	667.2	136
16	работать	611.2	148
17	понять	588.2	152
18	пойти	587.2	153
19	спросить	573.9	155
20	дать	573.1	157

Окончание ПРИЛОЖЕНИЯ 7

Пример части частотного словаря английских физических терминов

Материалом для частотного словаря послужили тексты по электронике, физике твердого тела и физике элементарных частиц на английском языке общим объемом 600 000 словоупотреблений. На первом этапе были получены три частотных словаря словоформ, которые затем были преобразованы в частотные словари лексем (исходных форм слов). Из этих последних извлечены списки терминов; будучи объединенными, они образовали алфавитно-частотный и частотный списки терминов.

i	*термины*	*F*
1	energy n	2168
2	value n	1897
3	temperature n	1691
4	electron n	1597
5	field n	1464
6	equation n	1452
7	fig. n	1432
8	obtain v	1351
9	state n	1340
10	current n	1260
11	function n	1236

i	*термины*	*F*
12	term n	1198
13	eq. n	1195
14	case n	1143
15	crystal n	1125
16	result n	1063
17	ion n	916
18	time n	899
19	effect n	870
20	region n	840
21	form n	833
22	particle n	829

Условные обозначения:

i – порядковый номер (ранг) слова в списке по убыванию частот,

F – частота слова в объединенном корпусе текстов.

ПРИЛОЖЕНИЕ 8

Пример словаря стоп-слов*

Данный словарь стоп-слов используется при индексировании и поиске документов БД по изобретениям и полезным моделям ФИПС (Федерального института промышленной собственности).

АНГЛИЙСКИЙ ЯЗЫК

A A ABOUT AFTER ALL ALSO AN AND ANY ARE AREN'T

B BECAUSE BEEN BEFORE BUT

C CAN CAN'T COULD COULDN'T

D DID DIDN'T DO DOES DOESN'T DON'T

E EACH EVEN

F FOR FROM

H HAD HADN'T HAS HASN'T HAVE HAVEN'T HE'LL HE'S HER HIM HIS HOW

I I'LL I'M ITS I'VE IF INTO ISN'T IT'S

J JUST

L LIKE

M MADE MAN MANY MAY ME MOST MR MUCH MUST MY

N NEW NOT NOW

O OF ON ONLY OR OTHER OUR OUT OVER

S SAID SHE SHE'LL SHE'S SHOULD SHOULDN'T SOME SUCH

T THAN THAT THE THEIR THEM THEN THERE THESE THEY THEY'RE THEY'VE THIS THOSE THROUGH TOO

W WAS WASN'T WAY WE WELL WE'RE WERE WE'VE WHAT WHEN WHERE WHICH WHO WITH WON'T WOULD WOULDN'T

Y YOU YOU'RE YOU'VE YOUR

*Открытый источник -

http://www1.fips.ru/wps/wcm/connect/content_ru/ru/inform_resources/inform_retrieval_system/article_4/article_7

ПРИЛОЖЕНИЕ 9

Список русско-английских терминов по курсу лекций

N	Термин на русском языке	Термин на английском языке
	Лекция 1	
1.	Лингвистика, Языкознание, Языковедение	*Linguistics*
2.	Фонетика	*Phonetics*
3.	Фонология	*Phonology*
4.	Лексикология	*Lexicology*
5.	Морфология	*Morphology*
6.	Синтаксис	*Syntax*
7.	Прагматика	*Pragmatics*
8.	Семантика	*Semantics*
9.	Дискурс	*Discourse*
10.	Общее языкознание	*General Linguistics*
11.	Описательное языкознание	*Descriptive Linguistics*
12.	Сравнительно-историческое языкознание, компаративистика	*Comparative Linguistics*
13.	Типология	*Linguistic Typology*
14.	Диалектология	*Dialectology*
15.	Когнитивная лингвистика	*Cognitive Linguistics*
16.	Паралингвистика	*Paralinguistics*
17.	Психолингвистика	*Psycholinguistics*, Psychology of language
18.	Социолингвистика	*Sociolinguistics*
19.	Прикладная лингвистика	*Applied linguistics*
	Лекция 2	
20.	Семиотика	*Semiotics*
21.	Семиология	*Semiology*
22.	Знак	*Sign*
23.	Знак-индекс	*Index sign*
24.	Знак-символ	*Symbol*
25.	Иконический знак	*Icon sign*
26.	Информатика	*Computer science*

27.	Компьютерная лингвистика	*Computational Linguistics*
28.	Математическая лингвистика	*Mathematical Linguistics*
29.	Лингвостатистика, Квантитативная лингвистика	*Quantitative Linguistics*
30.	Автоматическая обработка языка	*Natural Language Processing*
	Лекция 3	
31.	Искусственный язык	*Artificial language*
32.	Язык науки	*Language of Science*
33.	Язык программирования	*Programming language*
34.	Теория формальных языков	*Formal Language Theory*
35.	Метаязык	*Metalanguage*
36.	Формальная грамматика	*Formal grammar*
37.	Порождающая грамматика	*Generative grammar*
38.	Терминальный символ	*Terminal symbol*
39.	Нетерминальный символ	*Nonterminal symbol*
40.	БНФ-нотация, Форма Бэкуса-Наура	*Backus–Naur Form,* *Backus Normal Form*
41.	Синтаксическая диаграмма	*Syntax diagram,* *Railroad diagram*
	Лекция 4	
42.	Моделирование	*Modelling*
43.	Модель языка	*Language model*
44.	Лингвистическая модель	*Linguistic model*
45.	Искусственный интеллект	*Artificial Intelligence*
46.	Экспертная система	*Expert system*
47.	База знаний	*Knowledge Base*
48.	Язык представления знаний	*Knowledge Representation Language*
49.	Логика предикатов	*Predicate logic*
50.	Логика времени	*Temporal logic*
51.	Семантическая сеть	*Semantic net*
52.	Фрейм	*Frame*
53.	Сценарий	*Script*
54.	Правила продукций	*Production rules*

	Лекция 5	
55.	Информационный поиск	*Information Retrieval*
56.	Информационно-поисковые системы	*Information Retrieval Systems*
57.	Поисковая машина	*Search engine*
58.	Индекс	*Inverted index*
59.	Индексирование	*Search engine indexing*
60.	Ключевое слово	*Keyword*
61.	Релевантность	*Relevancy*
62.	Язык запросов	*Query Language*
63.	Полнота поиска	*Recall*
64.	Точность поиска	*Precision*
65.	Дескрипторный язык	*Controlled indexing language*
66.	Контролируемый словарь языка	*Controlled vocabulary*
67.	Информационно-поисковый тезаурус	*Thesaurus for Information Retrieval*
	Лекция 6	
68.	Лексикография	*Lexicography*
69.	Машинная лексикография	*Computational Lexicography*
70.	Терминология	*Terminology*
71.	Управление терминологией	*Terminology Management*
72.	Лемматизация, морфологическая нормализация	*Lemmatization*
73.	Терминография	*Terminography*
74.	Идеографический словарь	*Ideographic dictionary*
75.	Тезаурус	*Thesaurus*
76.	Глоссарий	*Glossary*
77.	Визуальный словарь	*Visual dictionary*
78.	Частотный словарь	*Frequency dictionary, Frequency word list*
	Лекция 7	
79.	Машинный перевод	*Machine Translation*
80.	Машинный перевод на базе примеров	*Example-based Machine Translation*

81.	Машинный перевод на базе лингвистических правил	*Rule-based Machine Translation*
82.	Статистический машинный перевод	*Statistical Machine Translation*
83.	Трансфер	*Transfer*
84.	ТМ-системы, Системы памяти переводов	*Translation Memory Systems*
85.	CAT-системы	*Computer-Aided-Translation tools*
86.	Машинное обучение, самообучение компьютера	*Machine learning*
87.	Перевод на базе облачных технологий	*Cloud-based translation*
88.	Компании-поставщики лингвистических услуг	*Language Service Provider, LSP*
89.	Переводческая компания	*Translation Service Provider, TSP*
Лекция 8		
90.	Гипертекст	*Hypertext*
91.	Язык гипертекстовой разметки документов	*HyperText Markup Language, HTML*
92.	Автоматическая обработка текста	*Text processing*
93.	Проверка орфографии	*Spellchecking*
94.	Оптическое распознавание печатных символов	*Optical Character Recognition, OCR*
95.	Распознавание рукописных символов	*Intelligent Character Recognition, ICR*
96.	Автоматическое реферирование	*Automatic Text Summarization*
97.	Анализ и синтез речи	*Automatic Speech Processing*
98.	Корпусная лингвистика	*Corpus Linguistics*
99.	Корпус текстов/ Корпусы текстов	*Corpus/* *Corpora*
100.	Компьютерная лингводидактика	*Computer Assisted Language Learning and Teaching, CALL, CALT*

СОДЕРЖАНИЕ

Учебное электронное издание

СОСНИНА Екатерина Петровна

ВВЕДЕНИЕ В ПРИКЛАДНУЮ ЛИНГВИСТИКУ

Учебное пособие

Усл. печ. л. 6,28.
Объем данных 1,33 Мб. ЭИ № 26.

Печатное издание
Изд. лиц. 020640 от 22.10.97
Подписано в печать 12.12.2112. Формат 60×84 / 16.
Усл. п. л. 6,28. Тираж 100 экз. (1-й з-д – 1–80 экз.) Заказ 1089.
Типография УлГТУ, 432027, Ульяновск, Сев. Венец, 32.

Ульяновский государственный технический университет
432027, г. Ульяновск, ул. Сев. Венец, 32.
Тел.: (8422) 778-113.
E-mail: venec@ulstu.ru