

Лекция 2. ЛИНГВИСТИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

План

1. Лингвистика: разделы и направления
2. Информационные технологии в лингвистике

1. Лингвистика: разделы и направления

Изучение возможностей применения информационных технологий в лингвистике предполагает знание основных понятий соответствующей области знания, среди которых можно выделить понятия из сферы лингвистики (язык, лингвистика, компьютерная лингвистика и т.п.) и информатики (информация, алгоритм, модель и др.).

Лингвистика (языкознание, языковедение) определяется как «наука о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях» (Лингвистический энциклопедический словарь)

Лингвистика (или языкознание) традиционно понимается как наука о естественном человеческом языке.

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

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

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

В.В. Звегинцев определил прикладную лингвистику как новую область лингвистики, «которая осуществляет реализацию лингвистических знаний с целью решения всякого рода практических задач».

В «Лингвистическом энциклопедическом словаре» прикладная лингвистика определяется как «направление в языкознании, занимающееся разработкой методов решения практических задач, связанных с использованием языка».

Задачи прикладной лингвистики делятся на традиционные, или «вечные», и «новые».

К числу традиционных задач относят:

- создание и совершенствование письменности,
- разработку систем транскрипции устной речи, систем транслитерации иноязычных слов,
- унификацию и стандартизацию научно-технической терминологии,
- создание словарей различных типов,
- перевод с языка на язык,
- обучение языку и т.д.

К новым задачам прикладной лингвистики относятся:

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

Основные направления прикладной лингвистики, связанные с изучением языка:

- ❖ лексикография — теория и практика составления словарей;
- ❖ лингводидактика — наука о разработках методик обучения иностранному языку;
- ❖ терминоведение — наука об упорядочении и стандартизации научно-технической терминологии;
- ❖ переводоведение — теория перевода.

Основные направления прикладной лингвистики (англ. Applied linguistics), связанные с практическими приложениями:

1) Компьютерная лингвистика (англ. Computational linguistics):

- Машинный перевод
- Автоматическое распознавание символов (англ. OCR)
- Автоматическое распознавание речи (англ. ASR)
- Автоматическое извлечение данных (англ. DataMining)
- Автоматическое реферирование текстов
- Построение систем управления знаниями
- Создание электронных словарей, тезаурусов, онтологий
- Корпусная лингвистика, создание и использование электронных корпусов текстов

2) Лингвистическая экспертиза (например, в судебной практике) наука об упорядочении и стандартизации научно-технической терминологии

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

Направления компьютерной лингвистики :

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

Язык в наиболее общем виде определяется как знаковая система, используемая для общения в некотором социуме. Различают естественные и искусственные языки.

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

Искусственные языки – знаковые системы, искусственно создаваемые в тех областях, где применение естественных языков менее эффективно или невозможно. Среди искусственных выделяются неспециализированные (или международные) языки (эсперанто, волапюк и др.) и специализированные языки: языки науки (математики, логики, химии и т.д.) и языки человеко-машинного общения (C++, Java, Python, Erlang и др.).

2 Информационные технологии в лингвистике

Информационные технологии в лингвистике – это совокупность законов, методов и средств получения, хранения, передачи, распространения, преобразования информации о языке и законах его функционирования с помощью компьютеров.

Задачи прикладной лингвистики:

- создание систем искусственного интеллекта;
- создание систем автоматического перевода;
- создание систем автоматического аннотирования и реферирования текстов;
- создание систем порождения текстов;
- создание систем обучения языку;
- создание систем понимания устной речи;
- создание систем генерации речи;
- создание автоматизированных информационно-поисковых систем;
- создание систем атрибуции и дешифровки анонимных и псевдоанонимных текстов;
- разработка различных баз данных (словарей, карточек, каталогов, реестров и т.п.) для гуманитарных наук;
- разработка различного типа автоматических словарей;
- разработка систем передачи информации в сети Интернет и т. д.

Эти комплексные задачи включают целый ряд более мелких проблем. К их числу относится автоматизация следующих процессов:

- построение словарей текстов;
- морфологический анализ слова;
- определение значения многозначного слова;
- синтаксический анализ предложения;
- поиск слова в словаре;
- порождение предложения и т. д.
- Одной из задач соответствующей области знания является сравнение способов кодирования информации человеком и компьютером.

Кодирование — это сопоставление объектов и отношений между ними с символами или словами какого-либо языка.

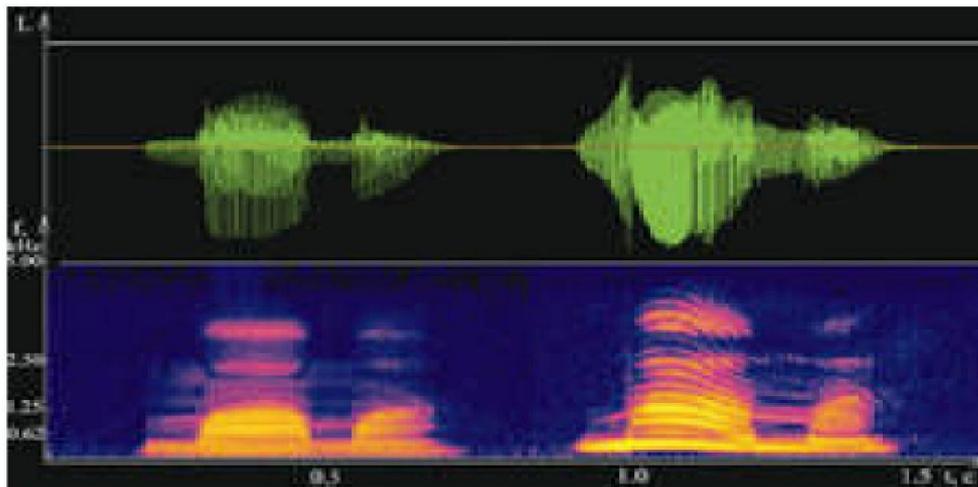
Компьютер может обрабатывать все известные виды информации

Для кодирования компьютером вербальной информации изначально использовался код ASCII (American Standard Code for Information Interchange).

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

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

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



Визуальное
представление
слова «мама»

Построение компьютерных лингвистических моделей предполагает выполнение некоторой последовательности действий. Формализованное описание такой последовательности действий, приводящей к решению поставленной задачи, называется алгоритмом.

Алгоритмы могут быть записаны в виде вербальных инструкций, блок-схем, таблиц или на языках программирования. Примеры алгоритмов различного рода см. в работе.

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

Thank you for your attention!