Компьютерная лингвистика

Лукашевич Наталья Валентиновна

louk_nat@mail.ru

План на сегодня

- Почему курс компьютерная лингвистика входит в программу факультета ИУ МГТУ им. Н.Э. Баумана

План на сегодня

- Почему курс компьютерная лингвистика входит в программу факультета ИУ МГТУ им. Н.Э. Баумана

- Будем обсуждать
  - Естественный язык и автоматическая обработка текстов
  - Информационный поиск

<u>Сегодня обзор задач и трудностей</u>

01-01. Автоматическая обработка текстов: Основные проблемы

Компьютерная лингвистика

- Текст-Речь-Диалог
- Автоматическая обработка текстов
- Natural language processing
- Human language technologies

- Искусственный интеллект

Естественный язык

- ЯЗЫК, система звуковых и письменных символов, используемых людьми для передачи их мыслей и чувств. (Кругосвет)

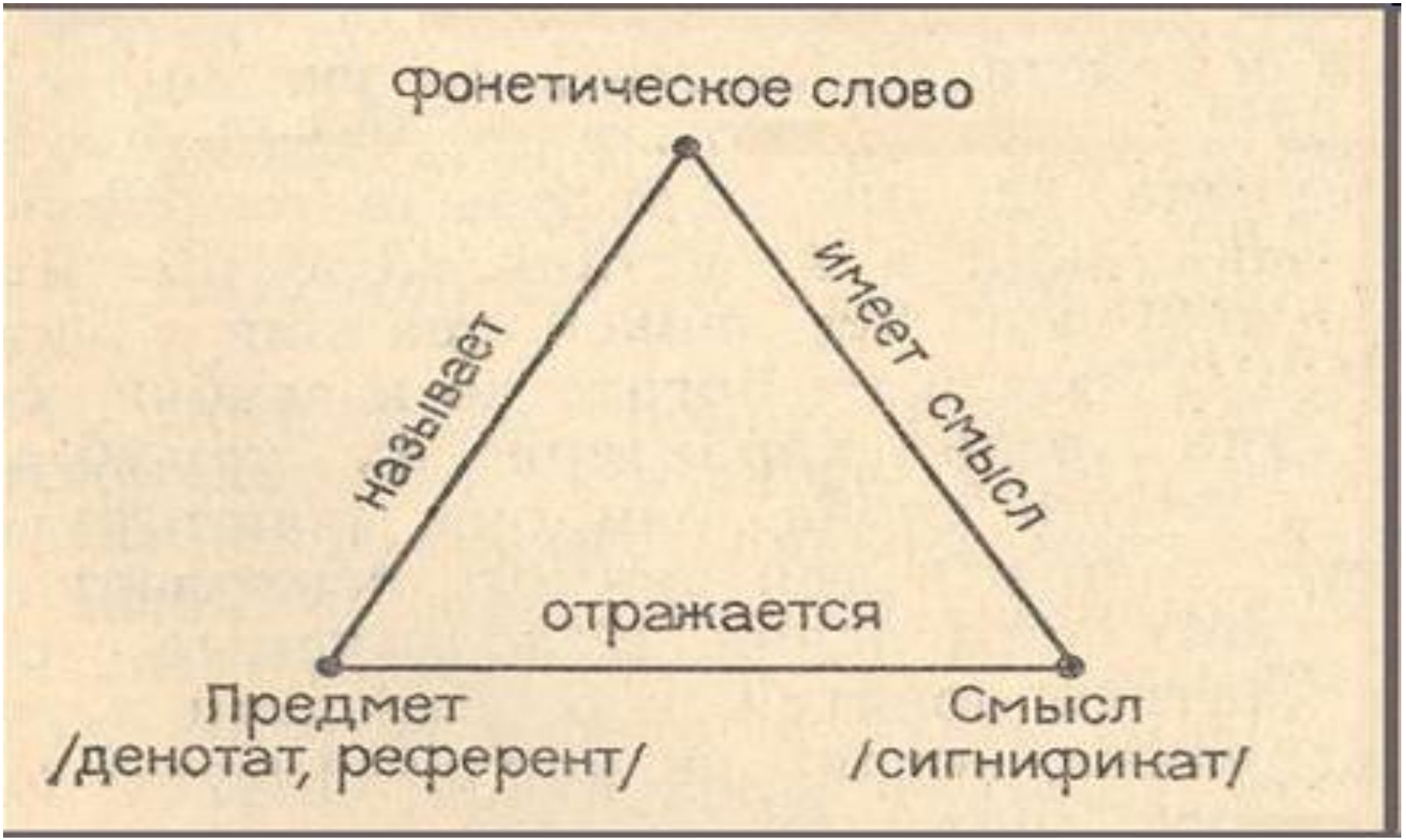
- Язык — знаковая система, используемая для целей коммуникации и познания

- Язык - исторически сложившаяся система словесного выражения мыслей, обладающая определенным звуковым, лексическим и грамматическим строем и служащая средством общения в человеческом обществе. (Словарь Ефремовой)

Языковой знак

- Означающее
  - последовательность звуков или графических знаков
- Денотат (референт)
  - обозначаемый предмет, явление действительности
- Означаемое (понятие)
  - отражение этого предмета, явления в сознании человека

Треугольник Фреге

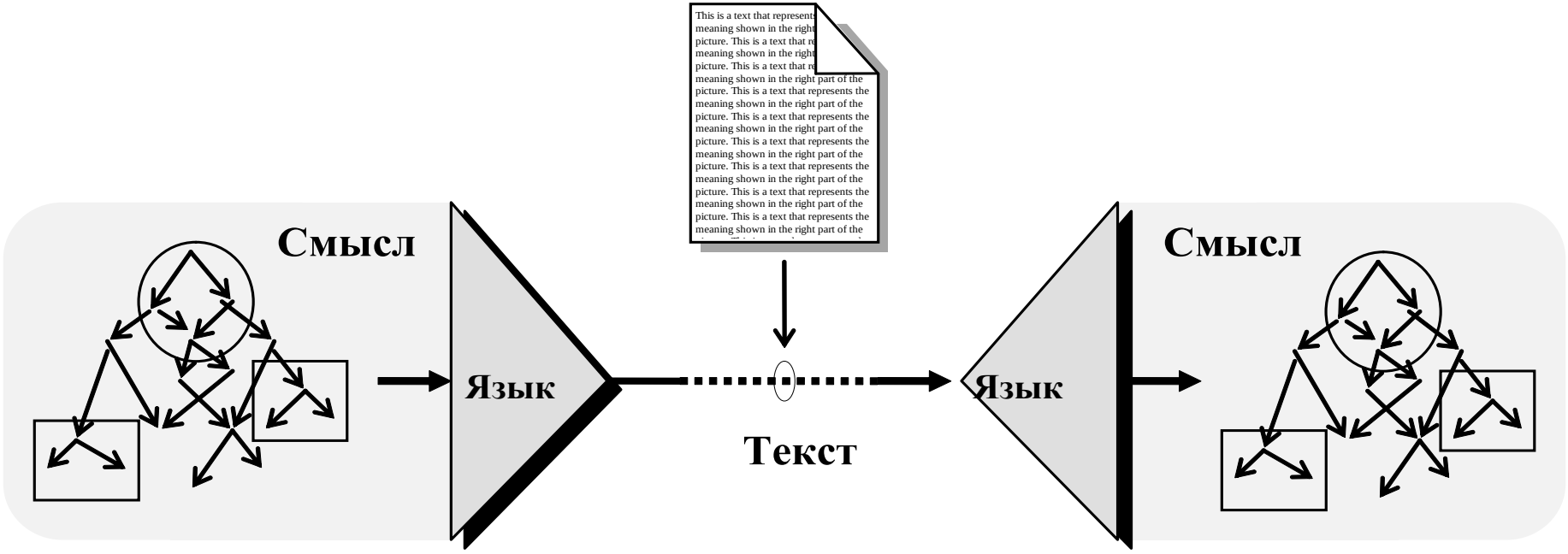


Функции языка

- Коммуникативные функции
  - констатирующая (для нейтрального сообщения о факте),
  - - информационная
  - вопросительная (для запроса о факте),
  - апеллятивная (для побуждения к действию),
  - экспрессивная (для выражения настроения и эмоций говорящего),
  - контактоустанавливающая (для создания и поддержания контакта между собеседниками);
- Функции мышления
- Функция хранения информации

Язык как преобразователь
Смысл ⇔ Текст

- Центральный объект – *текст,* линейность текста
- Текст составлен из различных *единиц*, относящихся к разным *уровням*
- Единицы: незначащие и значащие (языковые знаки)



Уровни языка

- Язык – иерархическая система
  - Фонема, (забор – запор)
  - Морфема (корень, приставка, суффикс, окончание)
  - Словосочетание
  - Предложение
  - Абзац
  - Текст

Уровни анализа текстов

- Графематический уровень
- Морфологический уровень
- Синтаксический уровень
- Семантический уровень
- Прагматический уровень

Основные свойства языка

- Многозначность
- Универсальность
- Избыточность
- Неопределенность и зависимость от контекста
- Изменчивость

Многозначность

- Фонетическая: лук (луг или лук)
- Морфологическая: стали
- Синтаксическая:
  - Девочка шла по полю с цветами
- Семантическая: *Возьми лук*

Неопределенность

- Город – поселок –деревня
- Река – ручей
- Высокий – низкий, молодой – старый

- Зависимость от контекста

Изменение языка во времени

- Изменчивость на протяжении поколения:
  - устаревание словарей,
  - появление новых слов

- Изменчивость языка лежит в основе образования семей и групп родственных языков.

Отличие естественных языков от языков программирования

- Открытость системы ЕЯ, изменчивость
- Нестандартная сочетаемость, частично описывается правилами грамматики (множество исключений) – heavy tea
- Многозначность на всех уровнях

Функциональные стили

- Разговорный (бытовой диалог)
- Литературно-художественный
- Газетно-публицистический (новостной)
- Научный
- Деловой (официально-деловой)

- Деловая проза – информативная функция языка

Уровни обработки текстов

Части речи

- Знаменательные части речи
  - Существительное (кто, что)
  - Глагол (что делать, что сделать)
  - Прилагательное (какой…)
  - Наречие (как, где…)
- Служебные части речи (закрытый список)
  - Междометия
  - Предлоги
  - Союзы
  - Частицы
  - Местоимения

Структура слова

- Части слова (морфемы)
  - Корень
  - Приставка
  - Суффикс
  - Окончание
- Операции со словами
  - Словоизменение
  - Словообразование

Словоизменение

- Изменение одного и того же слова по роду, числу, падежу
- Меняется окончание
- Лексема
- Нормальная форма слова vs. словоформы
- Часть речи не меняется
- Смысл слова не меняется

- Морфологические классы слов + правила словооизменения
- **Морфологический анализ**

Обработка словоформ:

Морфологический анализ

исследовать	{исследовать} + +Неопр.ф.
исследую	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 1 л.
исследуешь	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 2 л.
исследует	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 3 л.
…	
исследовал	{исследовать} + + Прош. вр. + Ед.ч. + М р.
исследовала	{исследовать} + + Прош. вр. + Ед.ч. + Ж р.
…	

Словообразование

- Образование новых слов
  - Корень-коренной,
  - пароход
- Может образоваться другая часть речи
- Смысл преобразуется
- Нет регулярности в образовании новых слов

Существительные

- Выражают основных действующих лиц
- Характеристики
  - Род
  - 3 склонения

- Словоизменение:
  - Число (множественное, единственное)
  - По падежам

Прилагательные

- Описывают свойства существительных
- Качественные и относительные
- Словоизменение:
  - Род
  - Число
  - Падеж
  - Краткая форма
  - Степени сравнения:
    - Сравнительная
    - Превосходная

Глаголы

- Описывают действия, состояния
- Характеристики
  - 2 спряжения
  - Вид глагола (совершенный, несовершенный)
- Словоизменение
  - Наклонение
  - Время
  - Лицо
  - Число
  - Формы глагола
    - Причастие (развитой – развивающийся)
    - Деепричастие

Местоимения

- Ссылаются на упомянутую сущность
- Осуществляют связь между частями предложения или между предложениями
- Разрешение референции
- Типы местоимений
  - Личные
  - Возвратное (себя)
  - Притяжательные (мой, твой, ваш)
  - Указательные местоимения
  - Вопросительные (относительные) местоимения

Другие части речи

- Предлоги
  - Однословные - многословные (в течение, в качестве)
  - Предлог определяет падеж стоящего после него слова
- Наречия
- Союзы

Словосочетания

– Устойчивые словосочетания, фразеологизмы
  - Не видно ни зги
  - Принять участие
– Свободные словосочетания (красивый дом)

  Словосочетания по главному слову
  - Именные группы (зеленый сад)
  - Предложные группы (в доме)
  - Глагольные группы (участвовать в игре)
  - Группа прилагательного (приятный на вид)
  - Причастный оборот (соответствующий образцу)
  - Деепричастный оборот (имея в кармане)

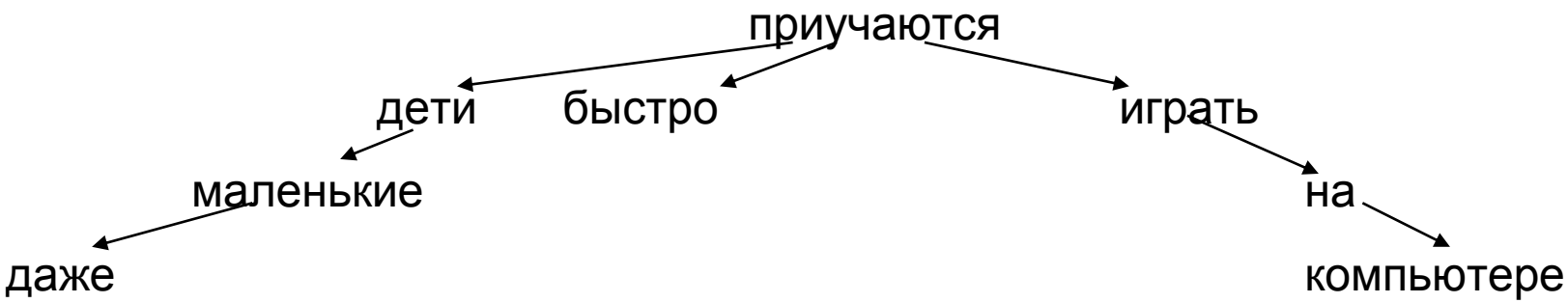
Отношения между словами в предложении (синтаксические связи)

- Согласование – род, число, падеж
  - Прилагательное согласуется с существительным
  - Подлежащее согласуется с глаголом

- Управление (аргументы, валентности)
  - Идти (кто, куда)
  - Купить (кто, что, у кого, за какую цену)
  - Семантические роли (агент, объект, место, время…)

Синтаксический анализ. Построение синтаксической структуры предложения

- Предложение рассматривается как конечное ***множество*** (элемент множества - словоупотребление).
- Всякое дерево, для которого данное предложение служит множеством узлов, называется ***деревом*** (синтаксического) ***подчинения*** для данного предложения.



Семантика

- Смысл слов, предложений, высказываний

- Хотелось бы
- - автоматически преобразовывать исходный текст в независимую от языка формальную запись. Но…

Лексическая семантика

- Многозначные слова
  - Критерии разделения слов на значения
- Лексические отношения
    - Синонимы
    - Антонимы,
    - Род-вид (гипероним – гипоним)
    - Часть-целое
    - Критерии установления лексических отношений
- Лексические функции:
  - Хороший (сон)=крепкий

Смысл предложения

- Проблема: смысл предложения трудно вывести из его частей
  - Пресуппозиции
    - Король умер=> Король существует
    - Король не умер=> Король существует
  - Следствия
  - Знания о мире
  - Идиомы - фразеологизмы

Прагматика

- Отношение говорящих к знакам
  - Явные и скрытые цели высказывания
    - *Выходите на следующей остановке?*
  - Максимы Грайса
  - Перформативы – высказывания, эквивалентные действию, поступку
    - Я клянусь
    - Я приказываю
  - Интерпретация слов типа *здесь, сейчас*

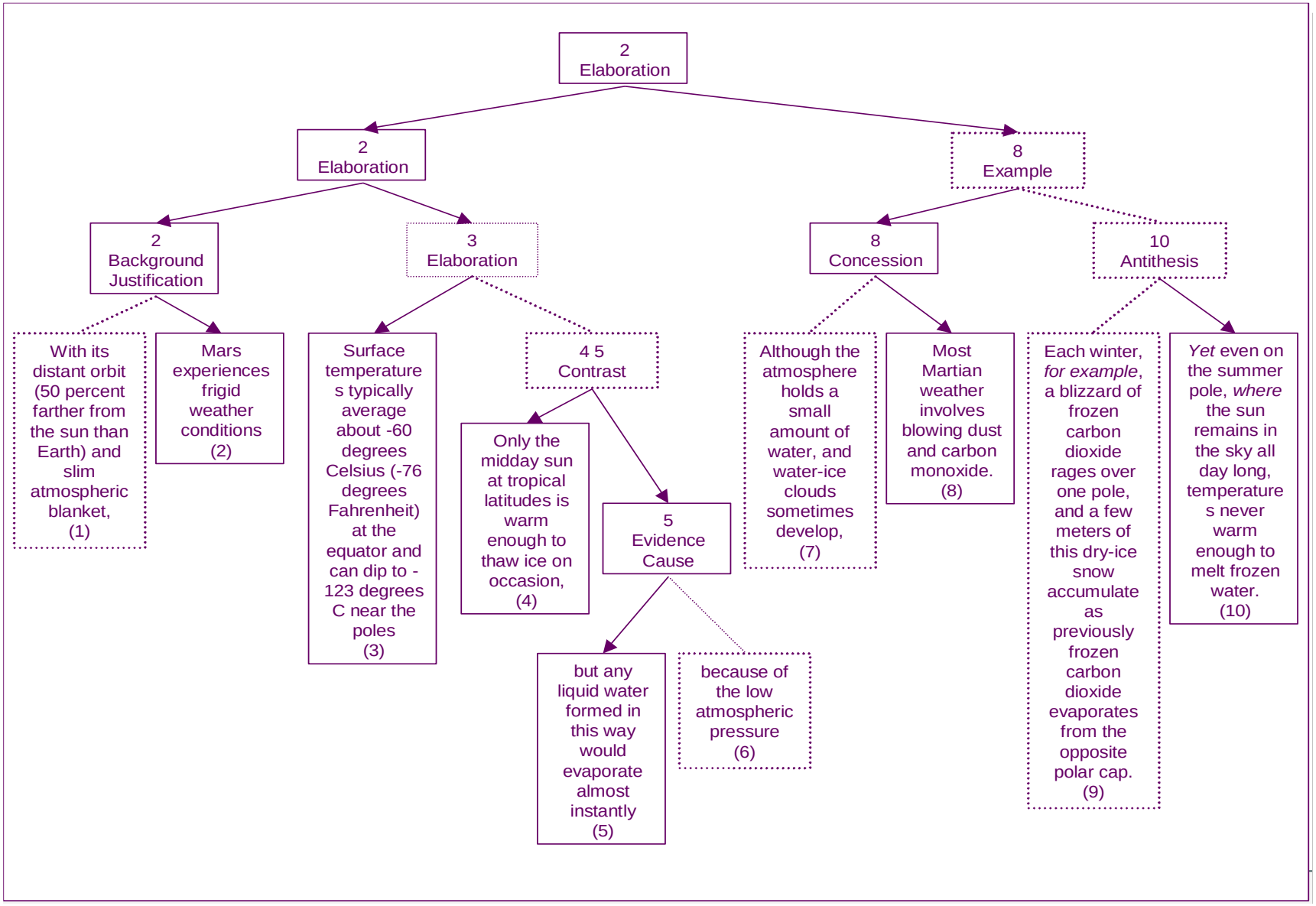
Смысл связного текста

- Неявная информация
- Излишняя информация
- Иерархическая структура текста
- Разные виды связей между предложениями
  - Дискурсивные слова
  - Анафорические связи
  - Повторы лексические и семантичские
  - Пропуски

Разрешение анафорических связей

- Сам **Евгений Чичваркин** приветствовал сегодняшний вердикт присяжных, в интервью радиостанции "Эхо Москвы" **он** заявил, что инициаторов этого процесса нужно судить
- ***Сбербанк*** *предупредил о возможных технических сбоях, теперь клиентам* ***банка*** *надо работать с банкоматами с особой осторожностью*

Представление связного текста в виде иерархической структуры



Лексическая связность текста

О порядке оказания безвозмездной финансовой помощи на строительство (покупку) жилья и выплаты денежной компенсации за наем (поднаем) жилых помещений военнослужащим и гражданам, уволенным с военной службы

Во исполнение Закона Российской Федерации "О статусе военнослужащих" и в целях обеспечения прав на жилище военнослужащих и граждан, уволенных с военной службы, Правительство Российской Федерации п о с т а н о в л я е т :

1. Утвердить прилагаемое Положение о порядке оказания безвозмездной финансовой помощи на строительство (покупку) жилья и выплаты денежной компенсации за наем (поднаем) жилых помещений военнослужащим и гражданам, уволенным с военной службы.

2. Министерству обороны Российской Федерации и иным федеральным органам исполнительной власти, в которых предусмотрена военная служба:

в месячный срок разработать и утвердить формы и перечень документов, необходимых для принятия решения об оказании военнослужащим безвозмездной финансовой помощи на строительство (покупку) жилья и о выплате денежной компенсации за наем (поднаем) жилых помещений;

расходы, связанные с оказанием военнослужащим безвозмездной финансовой помощи и выплатой денежной компенсации за наем (поднаем) жилых помещений, производить за счет и в пределах средств, выделяемых из федерального бюджета по сметам этих федеральных органов исполнительной власти.

3. Органам исполнительной власти субъектов Российской Федерации:

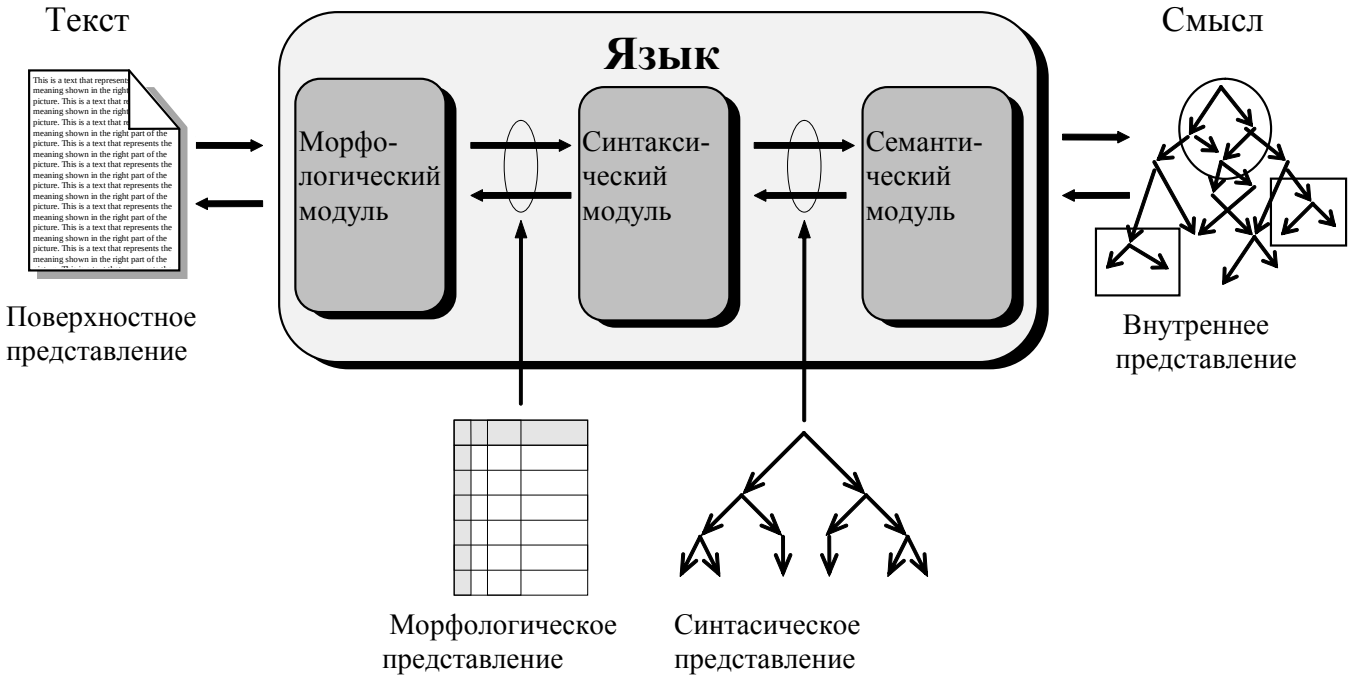
оказывать безвозмездную финансовую помощь в избранном постоянном месте жительства гражданам, уволенным с военной службы, осуществляющим строительство (покупку) жилья, за счет и в пределах средств федерального бюджета, выделяемых на жилищное строительство для этой категории граждан;

Уровни обработки текста

Сложность ЕЯ ⇒

лингвистический процессор – многоэтапный преобразователь

(два направления – анализ и синтез)



Лингвистические ресурсы

Лингвистические процессоры базируются на определенном представлении лингвистической информации:

- Компьютерные словари
- Грамматики ЕЯ
- Тезаурусы и онтологии

Источники лингвистической информации:

- Коллекции и корпуса текстов

Лингвистические ресурсы: словари и грамматики

- Компьютерные словари различаются:
  - Охватом лексики: общая/специальная
  - Представленной информацией:
    например: морфологические словари
  - Представленными единицами ЕЯ:
    - словари синонимов: *бродить / шататься*
    - словари паронимов: *чужой / чуждый*
    - словари терминов предметной области
    - словари (базы) устойчивых словосочетаний (*коллокаций*): *острая нехватка*, *задать вопрос*
- Грамматики – набор правил, описывающих синтаксическую структуру предложений:

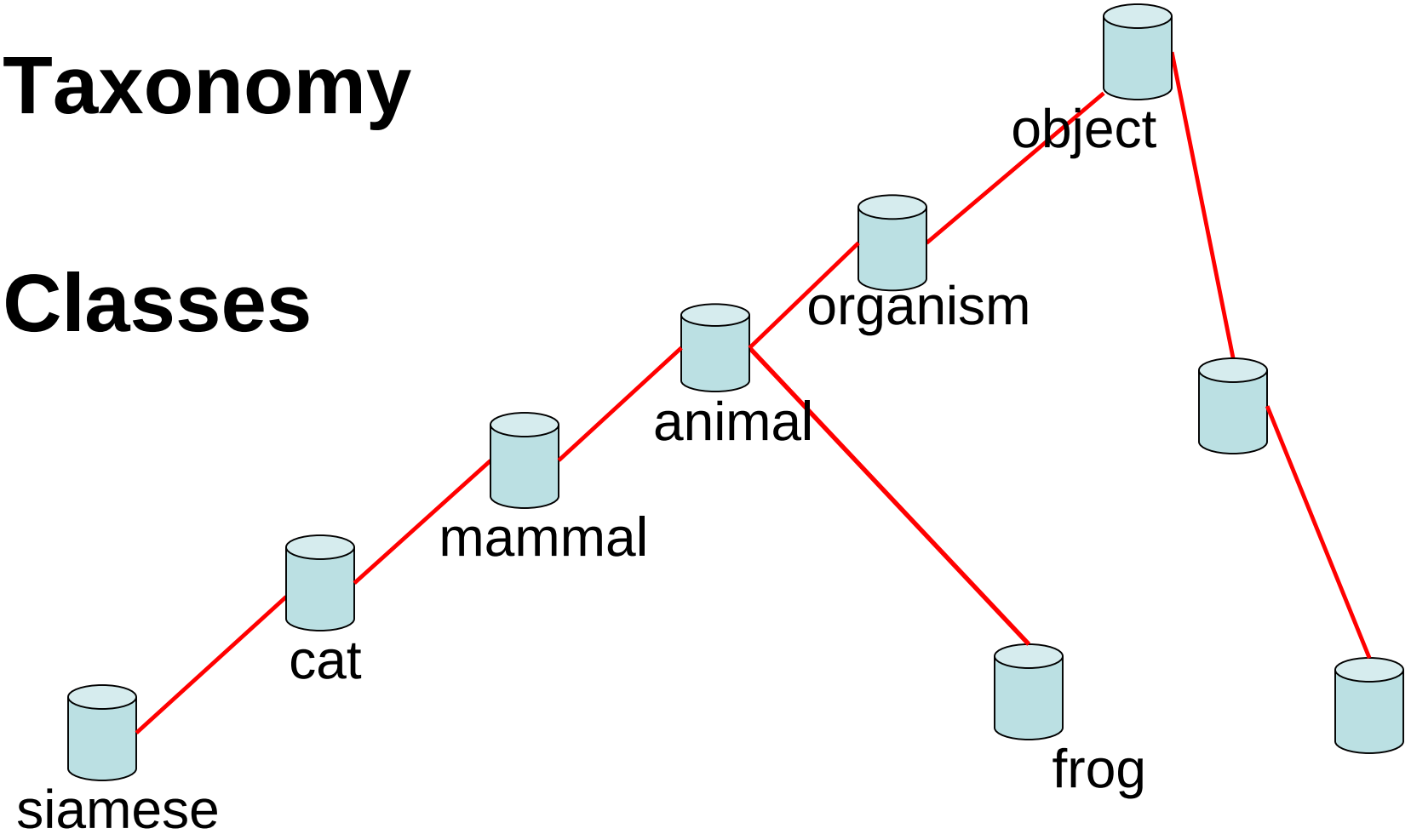
$$S \Rightarrow NP\ VP$$

Лингвистические ресурсы: тезаурусы и онтологии

- Тезаурус – семантический словарь
  - *РуТез* – информационно-поисковый тезаурус,
    - 52 тыс. понятий
    - связи: синонимия, род-вид, ассоциация, онтол. зависимость
- Онтология – формальное описание определенного набора понятий, сущностей
  - *WordNet* – лингвистическая онтология на базе англ. слов
    - Дж. Миллер (80е гг), модель человеческой памяти
    - слова разбиты по частям речи, для каждой части речи выделены *синсеты* (синонимы) - понятия
    - версия 3.0 – 155 тыс. лексем, 117 тыс. синсетов
  - *EuroWordNet* – аналогичные лексические ресурсы для других европейских языков

Taxonomy

Classes



instances

Лингвистические ресурсы: корпуса текстов

Трудоемкость создания
лингвистических процессоров и лексических ресурсов
$\Rightarrow$ автоматизация их построения

- *Коллекция текстов:* представительный набор текстов, собранных по определенному принципу
- *Корпус текстов:* коллекция текстов с *лингвистической разметкой*: морфологической, лексической, синтаксической, дискурсивной
  - использование в лингвистических исследованиях
  - применение для машинного обучения моделей

    для РЯ – Национальный корпус русского языка
- *Интернет-корпус:* тексты сети Интернет как корпус современной речи

Прикладные задачи автоматической обработки текстов

- Машинный перевод
- Извлечение информации из текстов
- Анализ мнений и оценка тональности текстов
- Генерация текстов на ЕЯ
- Автоматизация вёрстки и редактирования текстов
- Организация диалога на ЕЯ
- Обучение ЕЯ
- Распознавание и синтез звучащей речи
- Информационный поиск:
  - Классификация и кластеризация текстов
  - Реферирование и аннотирование текстов
  - Формирование ответов на вопросы