РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ ПРОБЛЕМ ПЕРЕДАЧИ ИНФОРМАЦИИ им. А.А.Харкевича

50 лет
1961
29 декабря
2011

Компьютерная лингвистика в ИППИ: история, современное состояние, перспективы

Ю.Д.Апресян, И.М.Богуславский, Л.Л.Иомдин

Лаборатория компьютерной лингвистики

Современное состояние

Фундаментальная научная задача: моделирование основных языковых способностей человека: понимания и говорения

Практическая реализация: многофункциональный лингвистический процессор ЭТАП-3

- многофункциональность означает применимость к любому классу задач, в которых в той или иной степени требуется понимание текстов и/или построение текстов по заданному смысловому представлению

Современное состояние: теоретические основы

И.А.Мельчук. Лингвистическая модель «Смысл ⇔ Текст»

Ю.Д.Апресян. Системная лексикография и теория интегрального описания языка

Современное состояние: основные опции лингвистического процессора ЭТАП-3

1. Машинный перевод
2. Анализ и синтез текстов на основе Универсального сетевого языка UNL
3. Синонимическое перифразирование
4. Семантический анализ текстов с участием онтологии
5. Гибридная система речевого синтеза русского текста
6. Лингвистическая разметка корпуса текстов
7. Синтаксический корректор

\+ Компьютерный учебник лексики

Современное состояние: основные особенности лингвистического процессора ЭТАП-3

Информатика

- Правиловый подход
- Самонастройка системы на обрабатываемый текст
- Ориентация на максимальное использование лингвистических ресурсов в разных приложениях (reusability)

Лингвистика

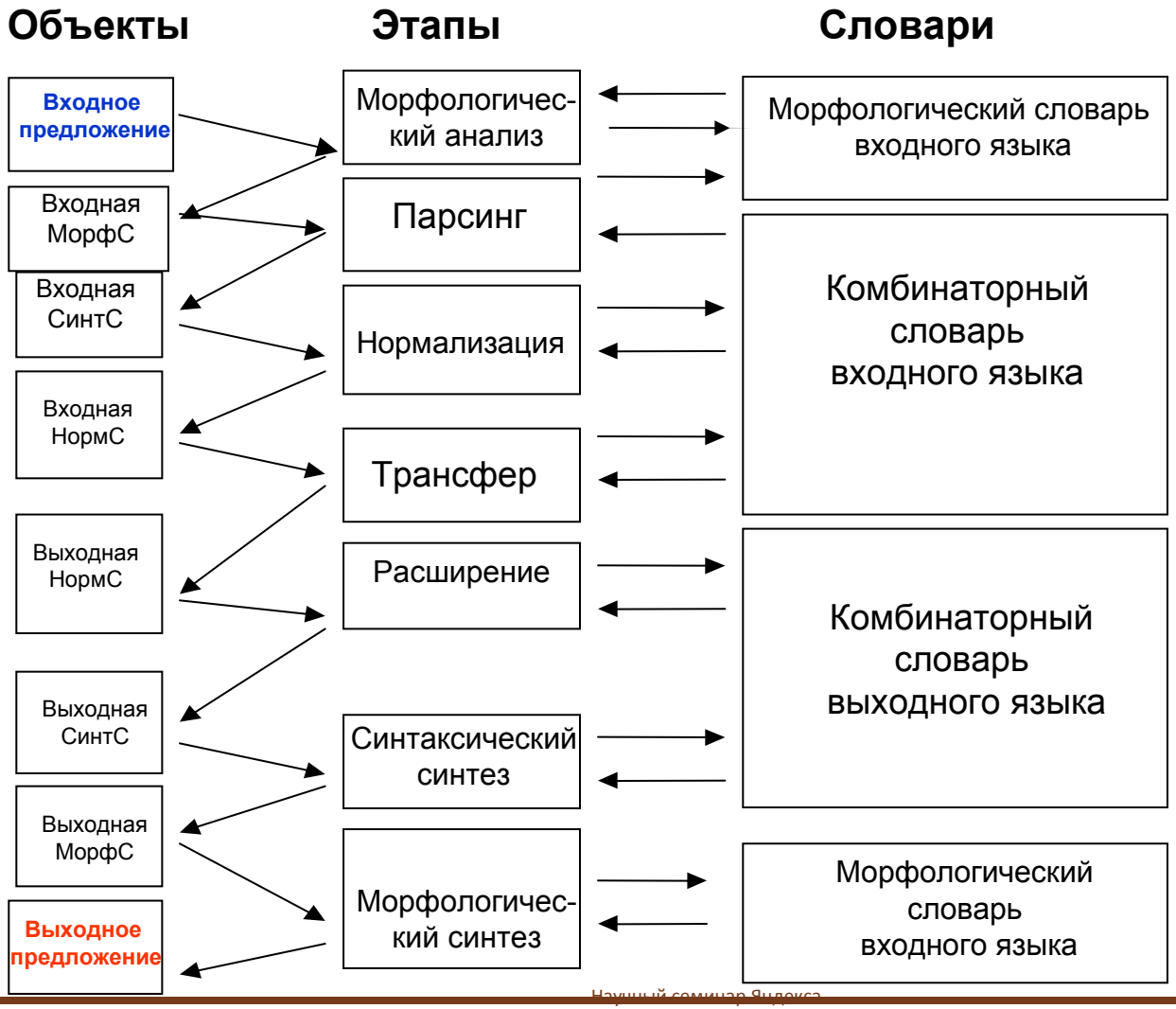
- Строгое разделение языковых уровней
- Дерево зависимостей как основной способ представления синтаксической структуры предложения
- Лексикалистский подход

Современное состояние: машинный перевод

- Русский ⇔ Английский
  - Морфологические словари объемом 130000 слов
  - Комбинаторные словари объемом 100,000 слов
- Русско-немецкий прототип
- Русско-французский прототип
- Русско-корейский прототип
- Испанско-английский прототип
- Англо-арабский прототип

Современное состояние: общая архитектура процесса перевода

07.12.2011



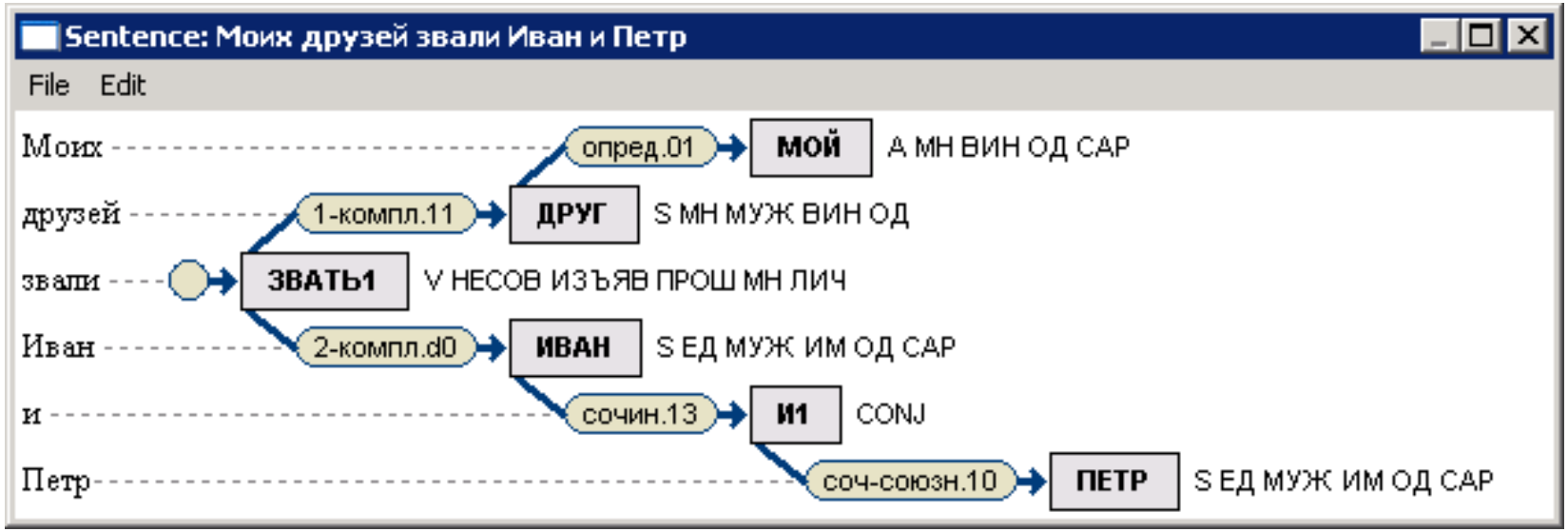
Современное состояние: машинный перевод

- Как переводятся неоднозначные предложения?

Моих друзей зовут Иван и Петр

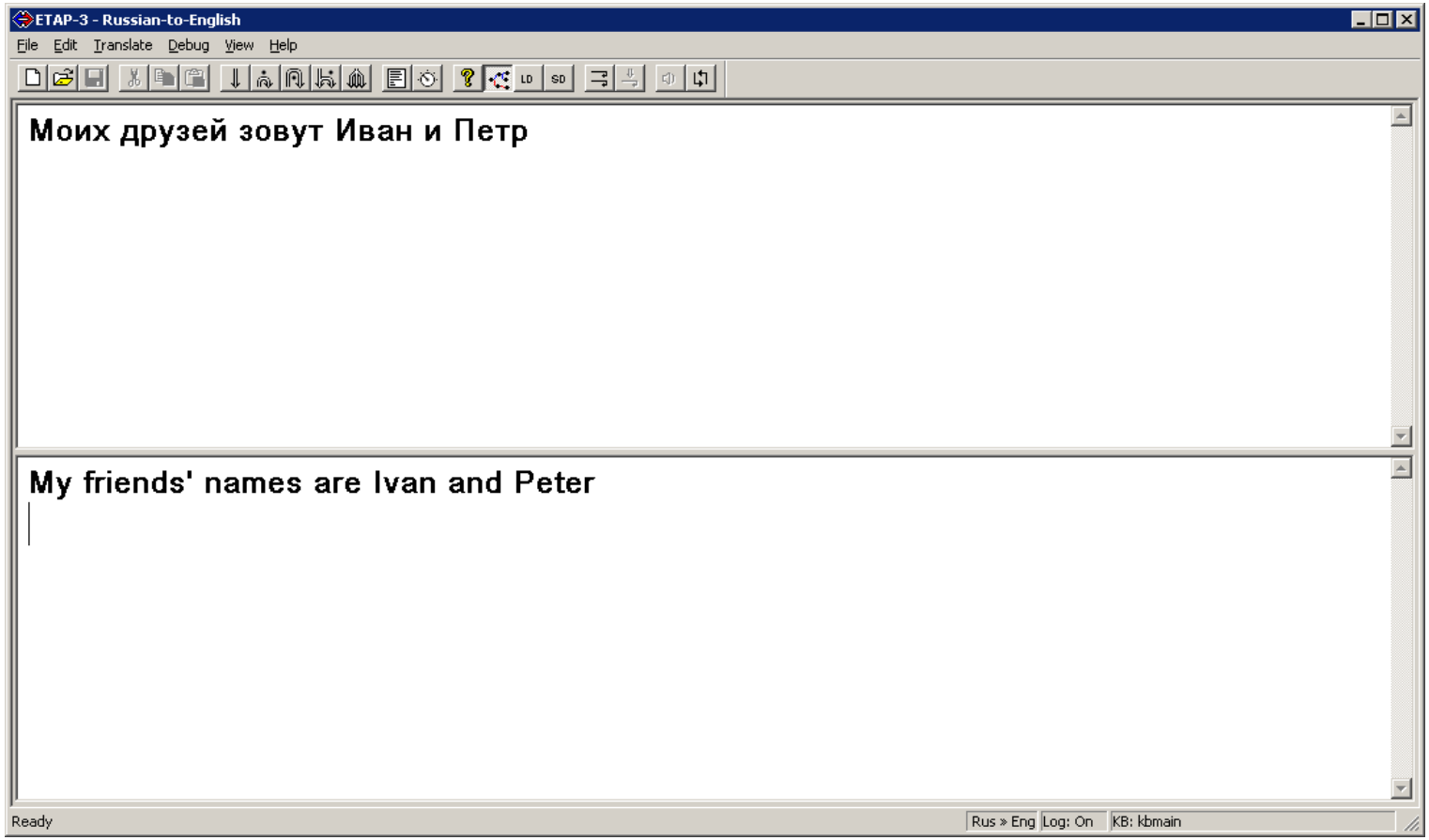
Современное состояние: машинный перевод

- Первый вариант древесной СинтС:



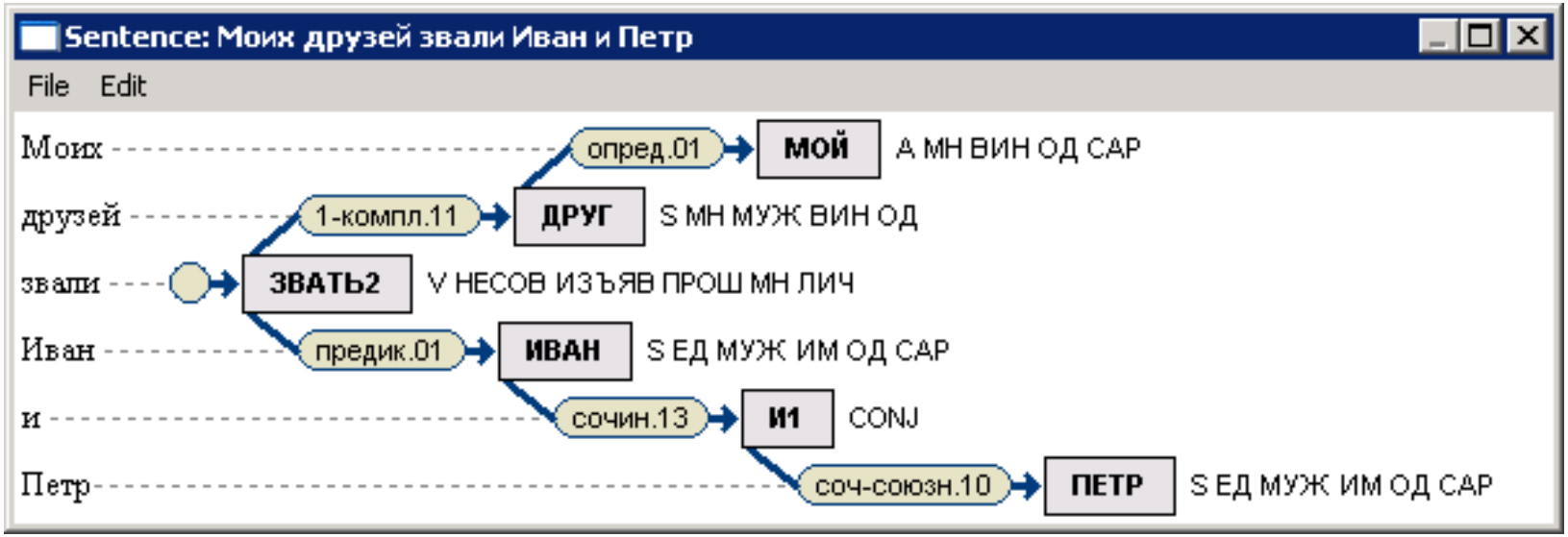
Современное состояние: машинный перевод

- Первый вариант перевода:



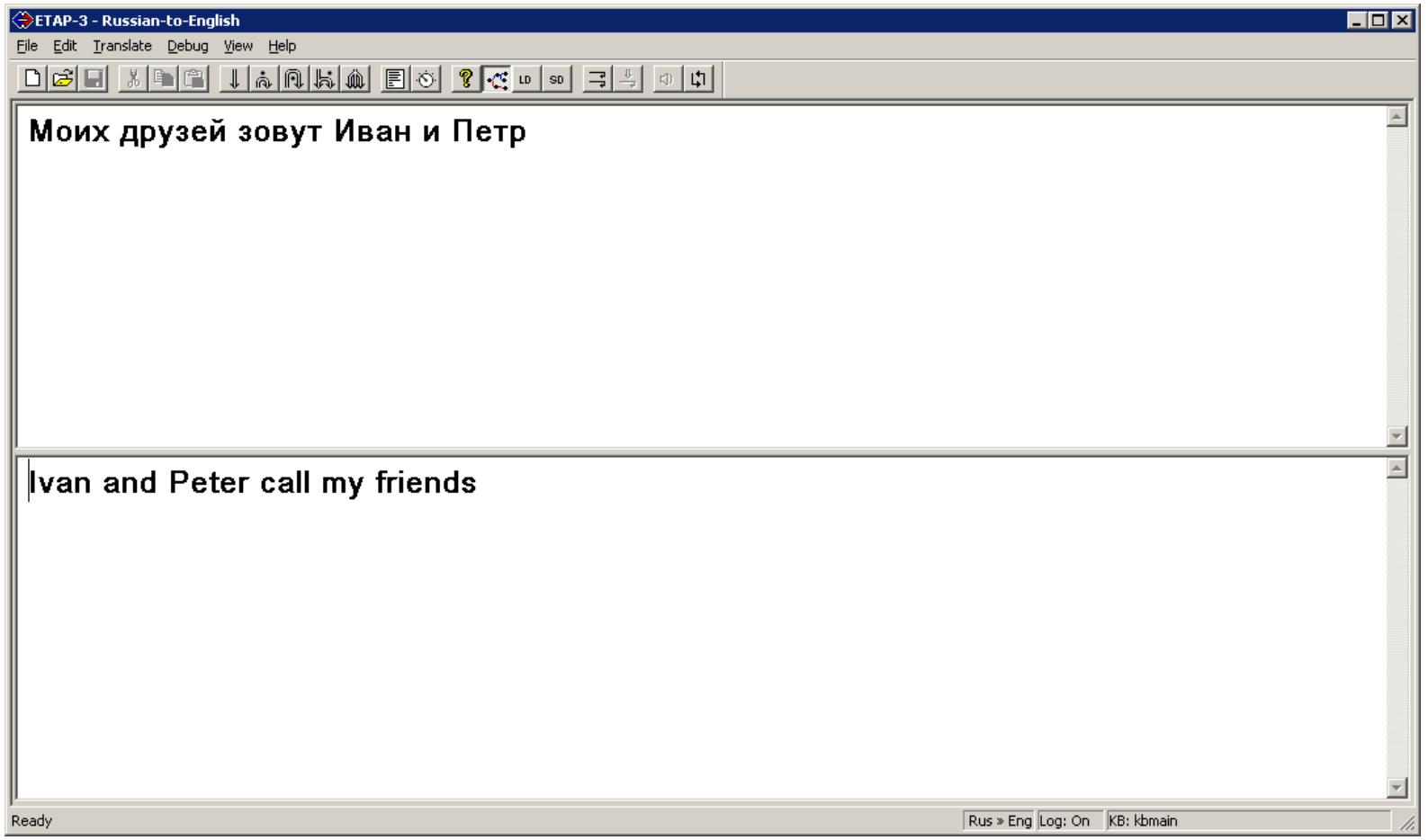
Современное состояние: машинный перевод

- Второй вариант древесной СинтС:



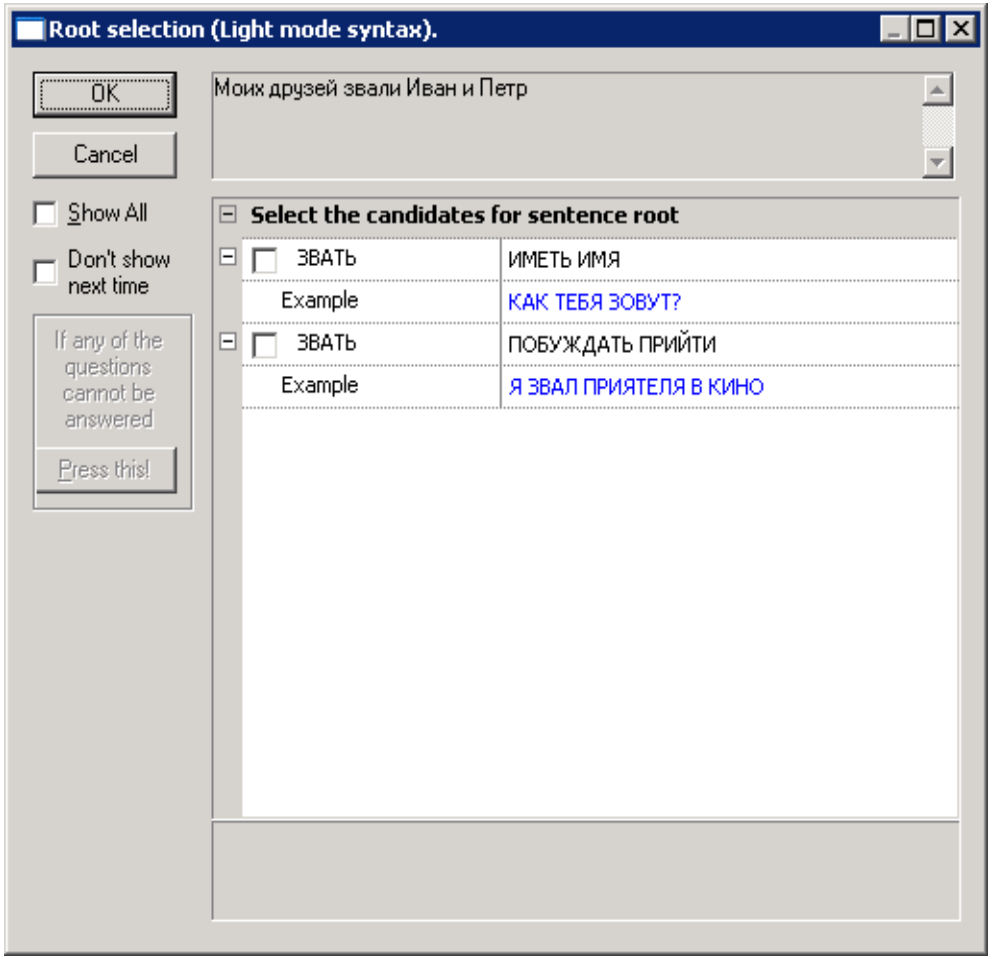
Современное состояние: машинный перевод

- Второй вариант перевода:



07.12.2011

Современное состояние: интерактивность в машинном переводе



Перспективы развития

1. Анализ текста
2. Синтез текста

Анализ текста. Фундаментальный аспект

- Переход на более глубокий уровень представления. Семантическая структура. СемС строится из толкований слов, грамматических категорий и других значимых единиц на основе механизма заполнения валентностей.
  - *Возмездие А1 человеку А2 за А3* ‘Действия человека А1, причиняющие серьёзный вред человеку А2, которые А1 совершает потому, что считает, что А2 сделал очень плохое А3 и за это должен испытать что-то, соразмерное А3 по степени вреда’
  - ПРОШ (Х) = ситуация Х имелп место до момента речи
- Подключение знаний о мире: онтология
- Логический вывод: извлечение из текста имплицитной информации.

Анализ текста. Прикладной аспект

Область, в которой существующие поисковые системы, основанные на статистике, не применимы: ответ на вопросы с использованием знаний о значениях слов, энциклопедической информации и логического вывода.

Сегодня, 14 ноября, дружина Петра Воробьёва в рамках чемпионата Молодёжной хоккейной лиги принимала на своём льду “Серебряных Львов” из Санкт-Петербурга. Единственную шайбу в матче забросил форвард ярославцев Максим Зюзякин.

Нужно ответить на вопрос, какие команды встречались, каков результат матча и где он проходил.

- *Сегодня, 14 ноября, дружина Петра Воробьёва в рамках чемпионата Молодёжной хоккейной лиги принимала на своём льду “Серебряных Львов” из Санкт-Петербурга. Единственную шайбу в матче забросил форвард ярославцев Максим Зюзякин.*

Языковые правила:

- команда, тренером которой является Пётр Воробьёв

Онтология:

Команда «Локо»

- **вид спорта: хоккей**
- **город: Ярославль**
- **главный тренер: Пётр Воробьёв**

- *Сегодня, 14 ноября, дружина Петра Воробьёва в рамках чемпионата Молодёжной хоккейной лиги принимала на своём льду “Серебряных Львов” из Санкт-Петербурга. Единственную шайбу в матче забросил форвард ярославцев Максим Зюзякин.*

- Матч проходил в Ярославле

- *Сегодня, 14 ноября, дружина Петра Воробьёва в рамках чемпионата Молодёжной хоккейной лиги принимала на своём льду “Серебряных Львов” из Санкт-Петербурга. Единственную шайбу в матче забросил форвард ярославцев Максим Зюзякин.*

- **Счёт в мачте – 1:0**

- *Сегодня, 14 ноября, дружина Петра Воробьёва в рамках чемпионата Молодёжной хоккейной лиги принимала на своём льду “Серебряных Львов” из Санкт-Петербурга. Единственную шайбу в матче забросил форвард ярославцев Максим Зюзякин.*

- **Победила команда из Ярославля, то есть «Локо»**

Синтез: от смысла к письменному тексту

Задел 1 (ИРЯ): интегральный словарь русского языка, содержащий всю информацию, которой надо владеть для правильного использования слова при говорении.

Задел 2 (ИППИ): словарь, насыщенный лексическими функциями.

Задел 3 (ИППИ): модуль синтеза текста ЭТАПа

Ожидаемый результат: грамматически правильный и идиоматичный текст, выражающий заданный смысл.

Синтез: от смысла к звучащей речи

- Синтезаторы речи
- Ошибки произношения в случае омонимии
  - ударение: *крУжка – кружкА*
  - произношение: *Все давным-давно надоели – Все давным-давно надоело; Все, что я видел – Все, кого я видел.*
- Неестественная интонация
- - синтаксическое строение предложения:

От радости // в зобу дыханье сперло

**От радости в зобу // дыханье сперло*

От боли в сердце // никто не застрахован

- - смысл: утверждение, вопрос, приказ, просьба...
- - информационная структура:

На улице много студентов (тема)

На улице // много студентов (контрастная тема).

Умный синтезатор

Задел 1: морфологический словарь ЭТАПа с информацией об акцентуации

Задел 2: морфологический и синтаксический анализатор ЭТАПа.

Задел 3: синтезатор речи (БАН)

Ожидаемый результат: синтезатор речи, способный вырабатывать правильную интонацию в зависимости от синтаксической, семантической и информационной структуры высказывания.