

Автоматический синтаксический анализ Введение

План

- Приложения NLP, требующие синтаксического анализа
- Подходы к представлению синтаксической структуры
- Примеры
- Лингвистические проблемы синтаксического анализа
- Формальные системы для реализации синтаксической разметки
- Алгоритмы для построения деревьев НС

План

- Приложения NLP, требующие синтаксического анализа. Зачем?
- Подходы к представлению синтаксической структуры
- Примеры
- Лингвистические проблемы синтаксического анализа
- Формальные системы для реализации синтаксической разметки
- Трибанки

Зачем

- Снятие морфологической омонимии
- Проверка правописания
- Вопросно-ответные системы
- Извлечение фактов
- Некоторые задачи автоматического реферирования
- Генерация текста
- Анализ мнений (opinion mining)
- Речевые технологии
- Создание и верификация лингвистических теорий (создание трибанков)

Зачем

- Борьба с неоднозначностью: морфология

| ЯРЛЫК СИТУАЦИИ | УСЛОВИЯ | ВЫВОД |
|--|--|---|
| Подсловарь СДС: "Тип омонимии $\langle A^*/N^* \rangle$ (<i>ego/ee/ix</i>)" \circ | | |
| <p>Ситуация n: \parallel</p> <p>$b_1 \dots \langle A^*/N^* \rangle \cdot (\text{ego/ee/ix}) \dots b_2$, где $b_1 \in B$ и $b_2 \in B$, $B = \{\text{тчк}, \text{тчк с зпт}, \text{двоет}, \langle PS^* \rangle, \langle Dv \rangle, \langle Vf \rangle, \langle Vinf \rangle, \langle Comp \rangle, \langle Npr \rangle\}$</p> | <p>между b_1 и b_2: $\exists \langle N^* \rangle \& \exists \langle A^* \rangle$</p> | $\langle N \rangle$ |
| Подсловарь СДС: "Тип омонимии $\langle A^*/N^* \rangle$ " \circ | | |
| <p>Ситуация $n+k$: \parallel</p> <p>$\langle P \rangle \cdot \langle A^*/N^* \rangle$</p> | <p>$P \cap (N^*_{\text{уир}} \langle A^*/N^* \rangle) = \emptyset$</p> <p>справа от $\langle A^*/N^* \rangle$ до $f \in N_j$: $P \cap N_j = \emptyset$</p> <p>$N_{\text{с12}} \cap (A^*_{\text{уир}} \langle A^*/N^* \rangle) = \emptyset$, где $f \in F$, $F = \{\text{тчк}, \text{тчк с зпт}, \text{двоет}, \langle PS^* \rangle, \langle Dv \rangle, \langle Vf \rangle, \langle Vsp \rangle, \langle Abr \rangle, \langle Avbr \rangle\}$</p> <p>справа от $\langle A^*/N^* \rangle$ до f: $\exists N_j \cdot P \cap N_j = \emptyset$</p> <p>$N_{\text{с12}} \cap (A^*_{\text{уир}} \langle A^*/N^* \rangle) = \emptyset$, где $f \in F$</p> | <p>$\langle A \rangle$</p> <p>$\langle N \rangle$</p> |

Зачем

- Борьба с неоднозначностью

- морфология:

*о довольной собой школьнице vs. К довольной собой школьнице vs. не встречал такой **довольной** собой школьницы;*

Мне грустно vs. Он грустно молчал vs. Это грустно

- лемматизация:

Три товарища vs. Три более тщательно

Глухой забор vs. Общество глухих

Зачем

- Проверка правописания:

границы оборотов, отделяемых запятыми

*Урсус жил в балагане на колесах, который Гомо, достаточно
вышколенный для этого, возил днем и стерег ночью.*

*Дом, который построил Свифт, — телевизионный художественный
фильм 1982 года режиссёра Марка Захарова.*

Зачем

Выделение терминологических словосочетаний (коллокаций)

- Он шел с [начальником отдела [*по* борьбе ...]]_{PP}]_{NP}
vs. Он [шел [*по* земле]]_{PP}]_{VP}
- [Уполномоченный президента [*в* Краснодарском крае]]_{PP}]_{NP} vs.
[[*В* краснодарском крае]]_{PP} произошло ...]

Зачем

- Границы именованных сущностей:
 - *[[Общество по защите прав потребителей] в Екатеринбурге] ОРГ*
 - *Грамотную защиту прав потребителей осуществит юридическая фирма «Манаенков и партнеры».*

Зачем

- Вопросно-ответные системы:
 - *Кто* был отправлен в командировку *Петровым*?
 - *Кто* отправил в командировку *Петрова*?

Зачем

Извлечение фактов:

- <http://viewer.opencalais.com/>

Извлечение мнений:

... неполной зум, большой экран, Хотелось бы оценить устройство зарядки и *прочность корпуса*. Недостатки: Макросъемка *оставляет желать лучшего*.

Зачем

- Извлечение онтологических знаний – автоматическое построение тезаурусов и онтологий с использованием лексико-грамматических шаблонов:
- «X это Y, который ...»
- X Y-а:
- *многоуровневая модель преобразований смысла в текст*

Зачем

- Алгоритмический машинный перевод:
 - *На месте пожара был обнаружен мертвым пожилый сторож*
 - *An elderly guard was found on the side of fire by the dead man*

(пример из презентации Б.Л.Иомдина)

Зачем

- Разметка корпусов – Трибанки
 - 1) собственно лингвистические исследования (частотность тех или иных конструкций, допустимость конструкций, ср. безвершинные относительные обороты, контроль деепричастных оборотов и т.п., виды эллипсиса)
 - 2) тренировка синтаксических анализаторов – технологии машинного обучения

План

- Приложения NLP, требующие синтаксического анализа. Зачем?
- Подходы к представлению синтаксической структуры
- Примеры
- Лингвистические проблемы синтаксического анализа
- Формальные системы для реализации синтаксической разметки
- Примеры алгоритмов: алгоритмы для построения деревьев НС

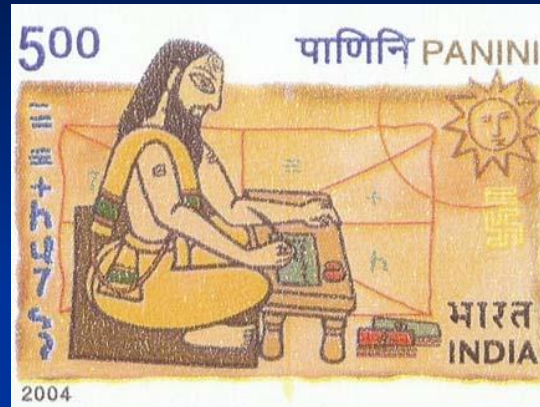
Формальное представление синтаксической структуры

- Морфологический анализ:
- *Партия скрипки*
- Синтаксический анализ:
 - синтаксический анализ - иерархичность
 - ??? каковы «следы» синтаксического анализа
 - ??? что является результатом синтаксического анализа?
 - *Онø увиде~~л~~ домø из песка* vs. *Онø увиде~~л~~ домø из окна*

Формальное представление синтаксической структуры

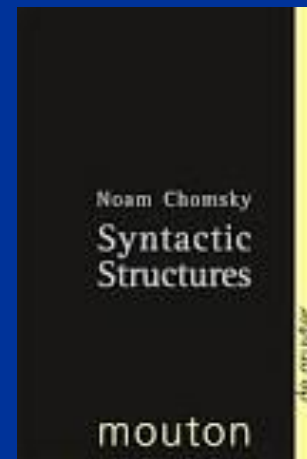
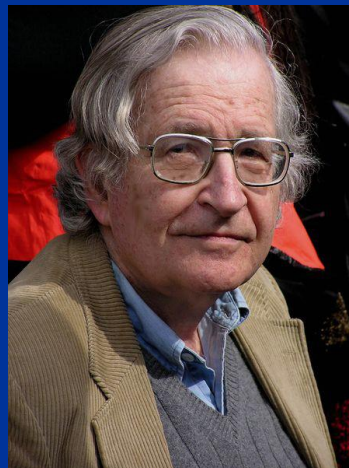
- определение единиц синтаксического описания
- фиксация синтаксических связей между этими единицами
- определения типа связи / функции синтаксического элемента

Представление синтаксической структуры



Панини
«Аштадхьяи»
(«Восьмикнижие»)

Ноам Хомский



Представление синтаксической структуры

- Грамматика зависимостей

Люсьен Теньер

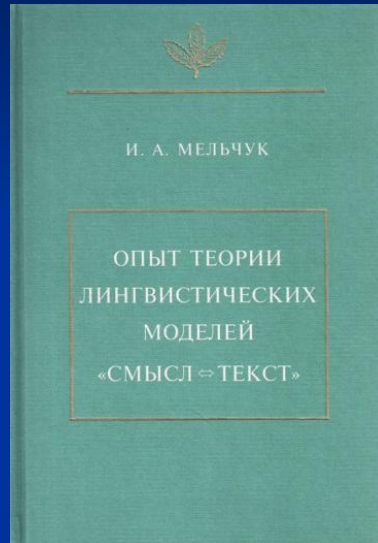


Представление синтаксической структуры

■ Грамматика зависимостей



Игорь Александрович
Мельчук
(российская традиция)



Paul M. Postal and David M.
Perlmutter. 1977.
Toward a Universal
Characterization of Passivization
(Stanford NLP)

Представление синтаксической структуры в системах АОТ

- грамматика зависимостей;
- грамматика непосредственных составляющих;
- традиционные синтаксические учения о членах предложения (корпус русского языка ХАНКО): грамматика структурных схем (Академическая грамматика-80);
- функциональная грамматика;
- семантический синтаксис (валентности, семантические роли);
- HPSG (Head-driven phrase structure grammar, грамматика управляемых вершинами фразовых категорий)
- категориальная грамматика
- грамматика связей (Link Grammar)
- и др.

План

- Приложения NLP, требующие синтаксического анализа. Зачем?
- Подходы к представлению синтаксической структуры
- Результаты синтаксического анализа. Примеры
- Лингвистические проблемы синтаксического анализа
- Формальные системы для реализации синтаксической разметки
- Трибанки

Примеры синтаксический разметки.

Пример 1. aot.ru



Примеры синтаксический разметки.

Пример 2. Грамматика зависимостей.

Connexor

Пример систем

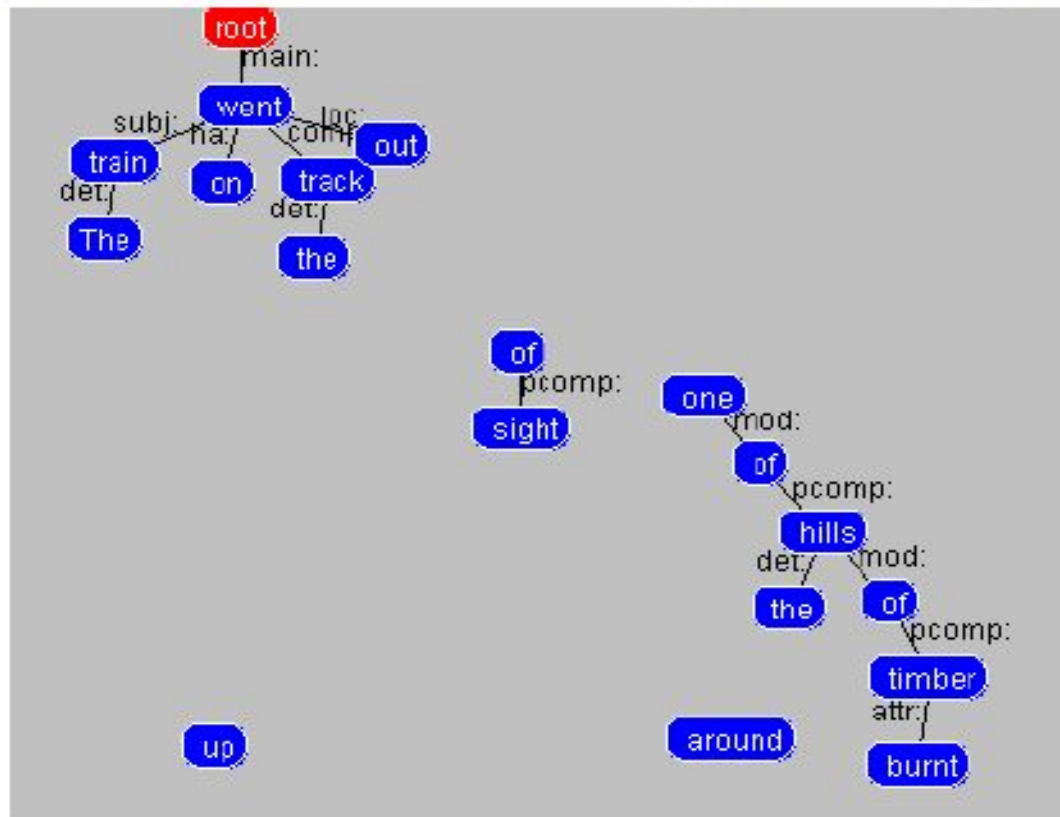
The train went on up the track out of sight, around one of the hills of burnt timber. Nick sat down on the bundle of canvas and bedding the baggage man had pitched out of the door of the baggage car.

Ernest Hemingway. Big two-hearted river

Примеры синтаксической разметки.

Грамматика зависимостей. Пример 2. Connexor

Connexor-1

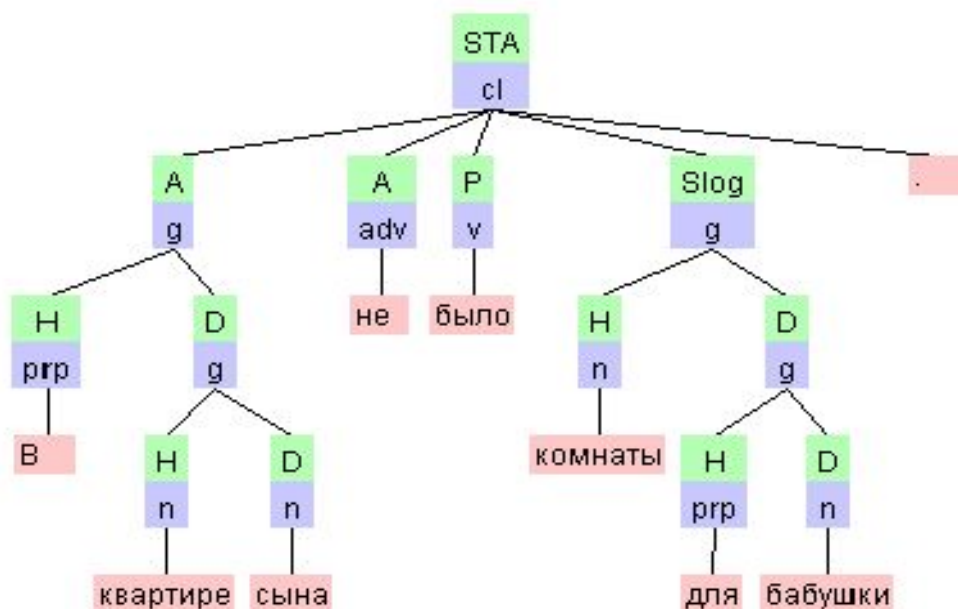


<http://www.connexor.com>

Примеры синтаксический разметки.

Пример 3. НС. Проект VISL

VISL



<http://visl.sdu.dk>

Примеры

Пример 4. Penn Treebank

tk treebank viewer

TREEBANK VIEWER Sandiway Fong University of Arizona (dec 2006: freeware version)

Sentence File /home/sandiway/treeprint/ws.j.txt Prolog Tree File /home/sandiway/treeprint/ws.j.pl Load

Sentence Count: 49208 Displayed Tree (Sentence): 1 Zoom: (x1) (x4)

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

S

NP-SBJ VP

NP ADJP MD VP

NNP NNP NP JJ old VB join NP PP-CLR NP-TMP

Pierre Vinken CD 61 NNS years will the board as DT JJ NN Nov. CD 29

a nonexecutive director

Компьютерная лингвистика. Гондова С.Ю.

Примеры

Пример 4. Penn Treebank

```
( (S
  NP Martin Marietta Corp.)
  was
  VP given
    (NP a
      $ 29.9
      million Air Force contract
      (PP for
        (NP low-altitude navigation
          and
          targeting equipment))))
.)
```

Примеры

Пример 5. Link Grammar Parser

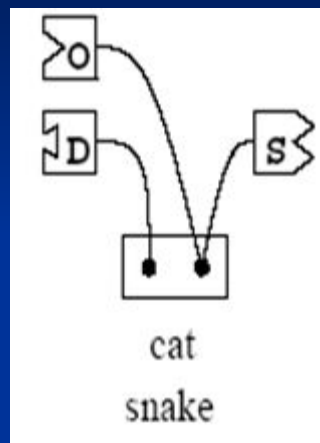


Рис.1

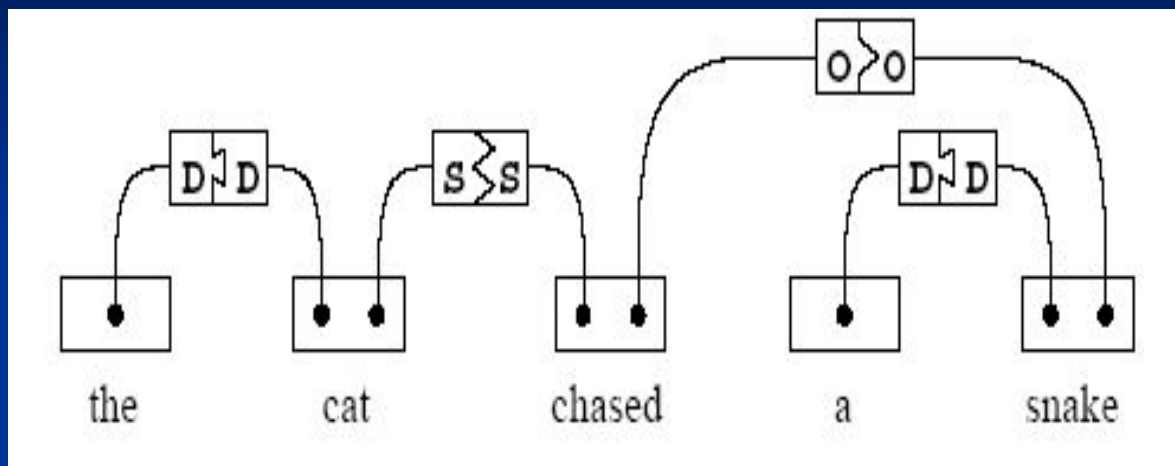
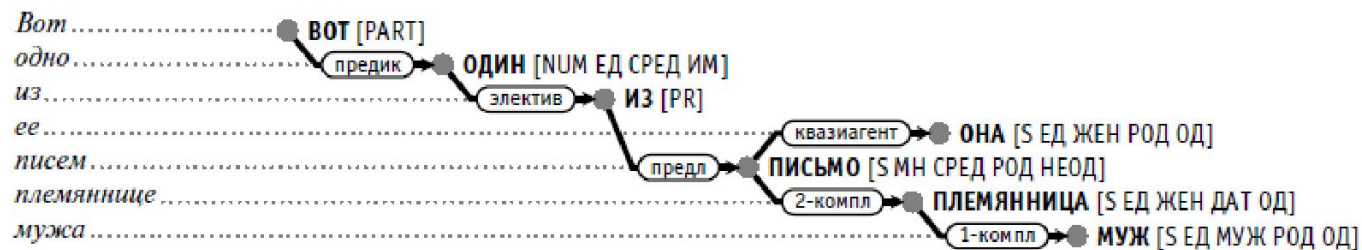


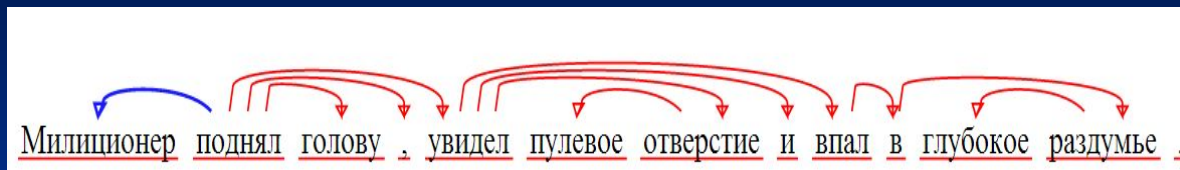
Рис.2

Все требования по связям слова описываются в словаре (рис.1). Прямоугольные блоки в виде разъемов — коннекторы. Коннектор может соединяться только с соответствующим ему коннектором с такой же формой разъема. Коннекторы, направленные направо, могут соединяться с коннекторами, направленными налево, и наоборот (рис.2)

Пример 5. СинТагРус



Пример 6. Rus-Treebank



SyntAutom

| id | token | type | head |
|----|---------------------|--------|------|
| 1 | Милиционер ← поднял | subj | 2 |
| 2 | поднял | fin | |
| 3 | голову ← поднял | acc | 2 |
| 4 | , ← поднял | conj | 2 |
| 5 | увидел ← поднял | homo | 2 |
| 6 | пулевое ← отверстие | adj | 7 |
| 7 | отверстие ← увидел | acc | 5 |
| 8 | и ← увидел | conj | 5 |
| 9 | впал ← увидел | homo | 5 |
| 10 | в ← впал | prepnp | 9 |
| 11 | глубокое ← раздумье | adj | 12 |
| 12 | раздумье ← в | prepnp | 10 |
| 13 | . | misc | |

План

- Приложения NLP, требующие синтаксического анализа. Зачем?
- Подходы к представлению синтаксической структуры
- Примеры
- Лингвистические проблемы синтаксического анализа
- Формальные системы для реализации синтаксической разметки
- Примеры алгоритмов: алгоритмы для построения деревьев НС

Лингвистические трудности

- 1) Синтаксическая омонимия

Торговка вяленой воблой торчала между ящиками

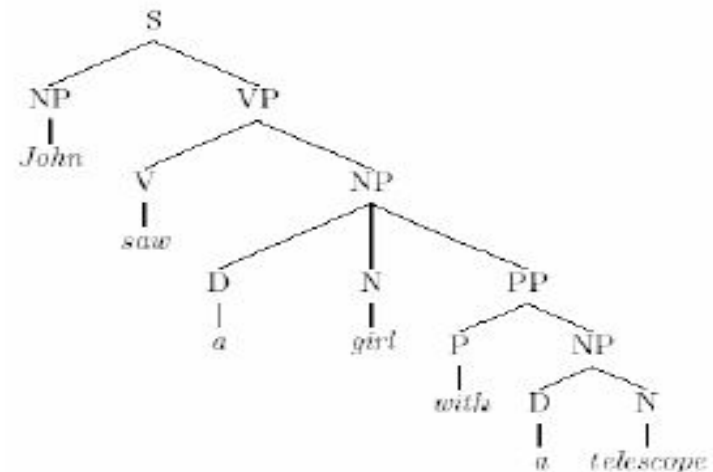
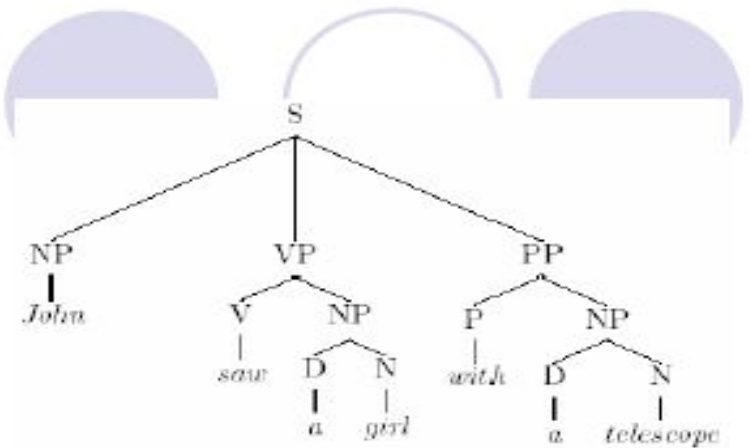
A diagram illustrating syntactic ambiguity. The sentence "Торговка вяленой воблой торчала между ящиками" is written in a cursive font. Two curved arrows originate from above the text: a teal arrow points to the word "Торговка" (the merchant), and a red arrow points to the word "торчала" (was sticking out).

Лингвистические трудности

Неоднозначности

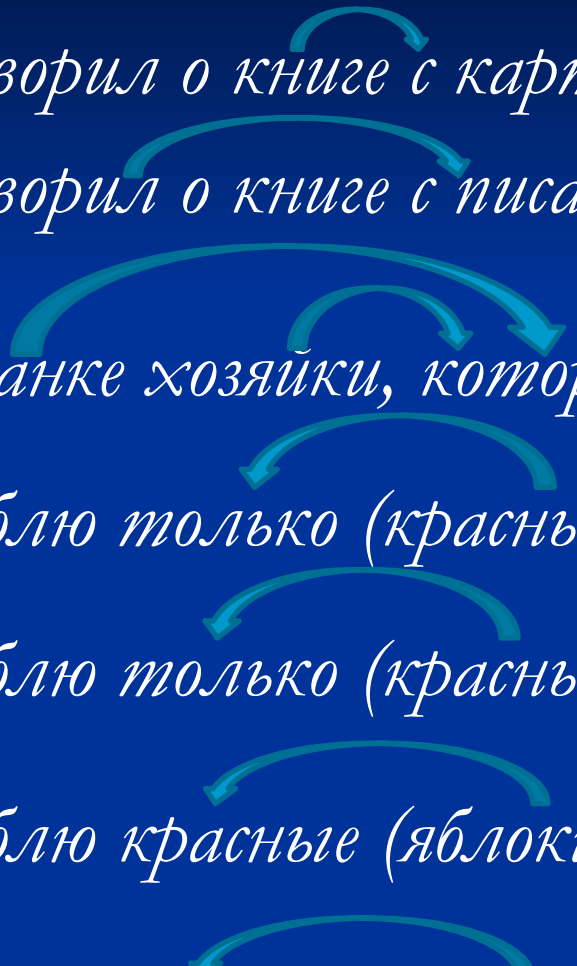
John saw a girl with a telescope.

Bertrand Gaie, Guy Perrier, 2004



Лингвистические трудности

Омонимия 1

- Он говорил о книге с картинками
 - Он говорил о книге с писателем
 - Служанке хозяйки, которая стояла на балконе
 - Я люблю только (красные яблоки)
 - Я люблю только (красные) яблоки
 - Я люблю красные (яблоки и груши)
 - Я люблю красные (яблоки) и груши
- 

Лингвистические трудности

Омонимия 2. Неразрешенная неоднозначность на морфологическом уровне

- *Мать любит дочь*
- *Такие типы **стали** есть в цехе*

Лингвистические трудности

Омонимия 3. Синтаксические метки

- *Мне нравятся фотографии **Головина** (приименная субъектная/приименная объектная связь)*
- *Головин сделал фотографии*
- *На фотографии изображен Головин*

Лингвистические трудности.

«Пустые узлы» 3

- Петя ~~ø~~/был очень аккуратный
- Петя обещал это сделать, когда ~~ø~~/он придет
- Он любит красные яблоки и ~~??ø~~ груши
- Петя любит красные яблоки, а Вася зеленые ~~ø~~
- Петя пошел купить ~~ø~~ хлеба
- Дай мне ~~я~~много хлеба

Лингвистические трудности

4. Порядок слов (непроективность, разрывы составляющих)

- *Он из Германии туманной привез учености плоды*

- *Посадил дед репку*


Лингвистические трудности ??? Единственность вершины

- *Петя и Вася пришли*
- *Петя рассказал, **как** он это сделал*

Проблемы синтаксического анализа

- составные лексические единицы (*буду писать*)
- морфологическая и синтаксическая омонимия (*типы стали, he saw a girl with telescope, портрет отца*)
- синтаксические нули (*Он сказал, что придет*)
- представимость в выбранном формализме:
 - непроективность — разрыв и перемещение составляющих (*посадил дед пенку*)
 - порядок объединения в синтаксические группы
 - требование единственности вершины (*он знает, что сказать*)
 - требование бинарности (*Петя дал ему книгу*)

Синтаксическое представления

Аспекты реализации

Аспекты реализации СА

- Словари (информация об индивидуальных единицах языка)
- Формальные правила
- Взаимодействие с соседними уровнями обработки (морфологический анализ, семантический анализ)

Формальные системы для реализации синтаксической разметки

- Грамматика конечных автоматов (Finite-State Transition Network)
- Контекстно-свободные грамматики (CFG)
- Расширенные сети переходов (ATN)
- Категориальные грамматики
- Унификационные грамматики (например, Link Grammar Parser, HPSG и др.)
- Tree Adjoining Grammar

Трибанки

- Penn Treebank
 - Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM)
 - Rus-Treebank
- <http://rus-treebank.maimbava.net/res01/rtb.php>
- СинТагРус

Дополнительный материал

- <http://www.sussex.ac.uk/Users/johnca/elsps.html> – сравнение парсеров, основанных на разных формализмах (Context-Free Grammar (CFG), Definite Clause Grammar (DCG), Head-Driven Phrase Structure Grammar (HPSG), Lexical Functional Grammar (LFG), Lexicalized Tree Adjoining Grammar (LTAG))
- <http://nlpub.ru> – синтаксический анализ
- Ножов И.М. Морфологическая и синтаксическая обработка текста (модели и программы). 2003. (электронная публикация диссертации <http://www.aot.ru/technology.html>)
- <http://www.scs.leeds.ac.uk/amalgam/amalgam/multi-parsed.html>