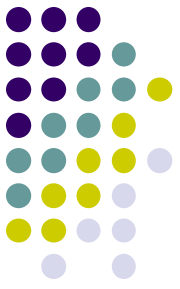




# СИНТАКСИЧЕСКИЙ АНАЛИЗ: МОДЕЛИ И ПРОЦЕССОРЫ

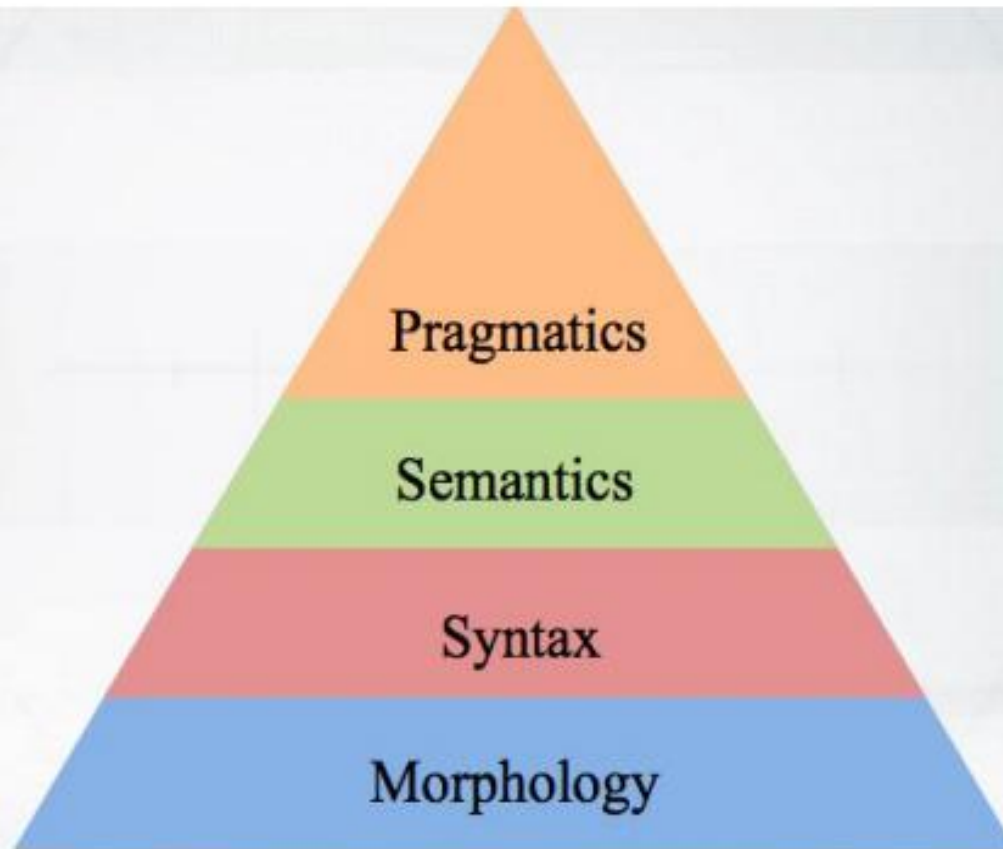
Большакова Елена Игоревна

# СОДЕРЖАНИЕ



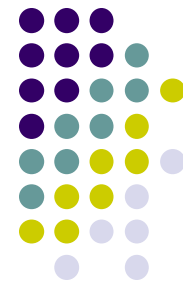
1. Основные модели синтаксич. анализа (СА)
  - структуры составляющих
  - структуры зависимостей:  
проективность, типы синтаксических связей
  - сравнение моделей, комбинированная модель
2. Синтаксические парсеры на основе правил
  - стратегии анализа, синтаксическая сегментация
  - парсеры *ЭТАП*, *Дуалинг-AOT*, *Compreno*
3. Синтаксические парсеры на базе обучения:  
*Stanford Parser*, *MaltParser*, *SyntaxNet*, *UDPipe*
4. Домашнее задание № 3

# УРОВНИ АНАЛИЗА ТЕКСТА



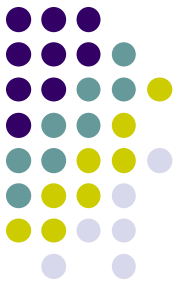
Natural Language Processing Pyramid

# ЭТАП СИНТАКСИЧЕСКОГО АНАЛИЗА



- Центральная роль синтаксиса: предложение выражает законченную мысль (предикативность):  
*This boy is tall. Она прочитала интересную книгу.*
- Единица обработки – *предложение*
- На входе: результат морфологического (и постморфологического) анализа  
На выходе: *синтаксическая структура* предложения
- Приложения синтаксического анализа:
  - Машинный перевод
  - Извлечение информации из текстов
  - Коррекция текстов на ЕЯ:  
исправление грамматических ошибок
  - Аннотирование текста (глубокое)
  - Вопросно-ответные системы
  - Обучение иностранным языкам

# ОСНОВНЫЕ МОДЕЛИ СА



- Сложность формализации и моделирования СА:
  - разные структурные особенности ЕЯ
  - кроме уровня предложений есть значимый подуровень **словосочетаний** (средство номинации)
- Модели СА отличаются в основном:
  - *синтаксическими единицами*
  - *синтаксическими связями между ними*
- Общее: *синтаксическое дерево* предложения – *Почему?*
- Основные подходы к моделированию синтаксической структуры предложений:
  - *структуры (деревья) зависимостей/подчинения*
  - *системы (деревья) составляющих*
- в лингвистике РЯ: *теория членов предложения*

# МОДЕЛЬ СИНТАКСИСА: ЧЛЕНЫ ПРЕДЛОЖЕНИЯ



- Отечественная лингвистика (русистика):  
теория членов предложения (изучается в школе)
- Синтаксические единицы – функциональные единицы:
  - *сказуемое*
  - *подлежащее*
  - *определение*
  - *дополнение (прямое/косвенное)*
- Два типа связи:
  - *взаимоподчинительная* (подлежащее – сказуемое)
  - *подчинительная* (образует иерархию членов):  
главные члены и второстепенные члены
- Нечеткость (непригодность) лингвистической модели  
для ее полной формализации

# МОДЕЛЬ СИНТАКСИСА: СТРУКТУРЫ ЗАВИСИМОСТЕЙ



- Восходит к Л.Теньеру (французский лингвист), существенное развитие – А. Реформатский, И. Мельчук
- Постулируется только **подчинительная связь**, **корень** синтаксического дерева – глагол (сказуемое)
- Синтаксические единицы (≈ члены предложения):  
*актанты , сирконстанты*
- Иерархия актантов по обязательности  
*субъект* (≈ подлежащее) – *объект*
- Определения подчиняются актантам и друг другу
- Синтаксические связи (разная степень дифференциации):
  - *опредетельное: желтое солнце*
  - *прямообъектное: читаю книгу*
  - *сочинительное: война и мир*
  - ....

# МОДЕЛЬ СИНТАКСИСА: СИСТЕМЫ СОСТАВЛЯЮЩИХ



- Возникла в американской лингвистике:  
метод непосредственно составляющих
- Синтаксические единицы – **составляющие** (*constituents*), т.е. отрезки текста (группы соседних слов), получающиеся в результате линейного членения предложения
- Могут вкладываться друг в друга:  
любые две составляющие либо не пересекаются,  
либо одна целиком содержится в другой
- Два крайних случая составляющих: слово, предложение
- Связь этих синтаксических единиц:  
ненаправленное **отношение вложения**  
*((Серая птичка) (весело (поет (незатейливую песенку))))*
- Фактически выделяются словосочетания (фразы - **phrases**) разных уровней (без иерархии фраз одного уровня)



# СИСТЕМА СОСТАВЛЯЮЩИХ: ФОРМАЛИЗАЦИЯ



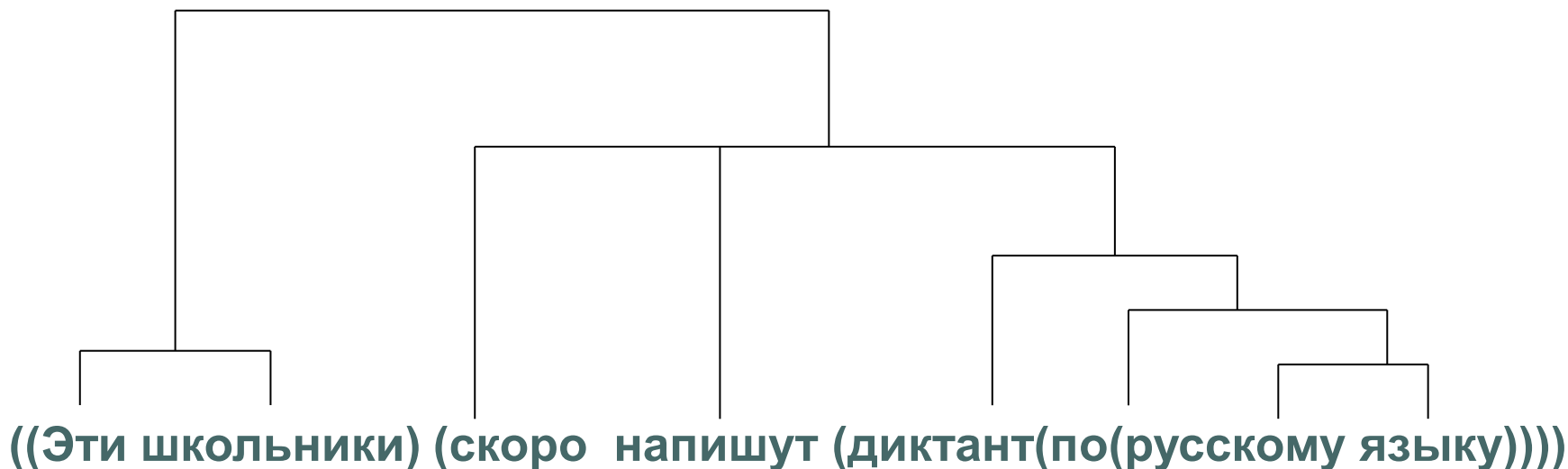
- Предложение – **цепочка** словоформ  
$$S = (w_1, w_2, \dots, w_N)$$

т.е. конечное линейное упорядоченное множество.
- *Составляющая* – произвольная подпоследовательность (*отрезок*) цепочки.
- *Система составляющих* – такое множество  $C$  отрезков этого множества  $S$ , которое удовлетворяет условиям:
  1.  $\forall w \in S : w \in C$  (т.е. слово – составляющая)
  2.  $S \in C$  (т.е. само предложение является элементом системы своих составляющих)
  3.  $\forall \alpha, \beta$ , таких что  $\alpha \in S, \beta \in S$   
и либо  $\alpha \cap \beta = \emptyset$ , либо  $\alpha \subset \beta$ , либо  $\beta \subset \alpha$   
(т.е. любые две составляющие или не пересекаются, или одна из них вложена в другую)

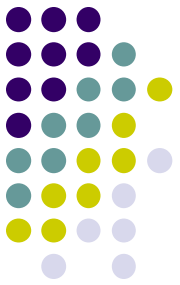
# ДЕРЕВЬЯ СОСТАВЛЯЮЩИХ



- Вложенность отрезков системы составляющих  $S$  изображается как *дерево составляющих*
  - листья – слова предложения
  - поддеревья – фразы (составляющие)
  - дуги – отношения вложения
- Для предложения  $S$  возможны разные деревья/системы  $S$  среди них только некоторые отражают принятые в лингвистике соглашения о граммат. структуре предложения



# РАЗМЕЧЕННАЯ СИСТЕМА СОСТАВЛЯЮЩИХ

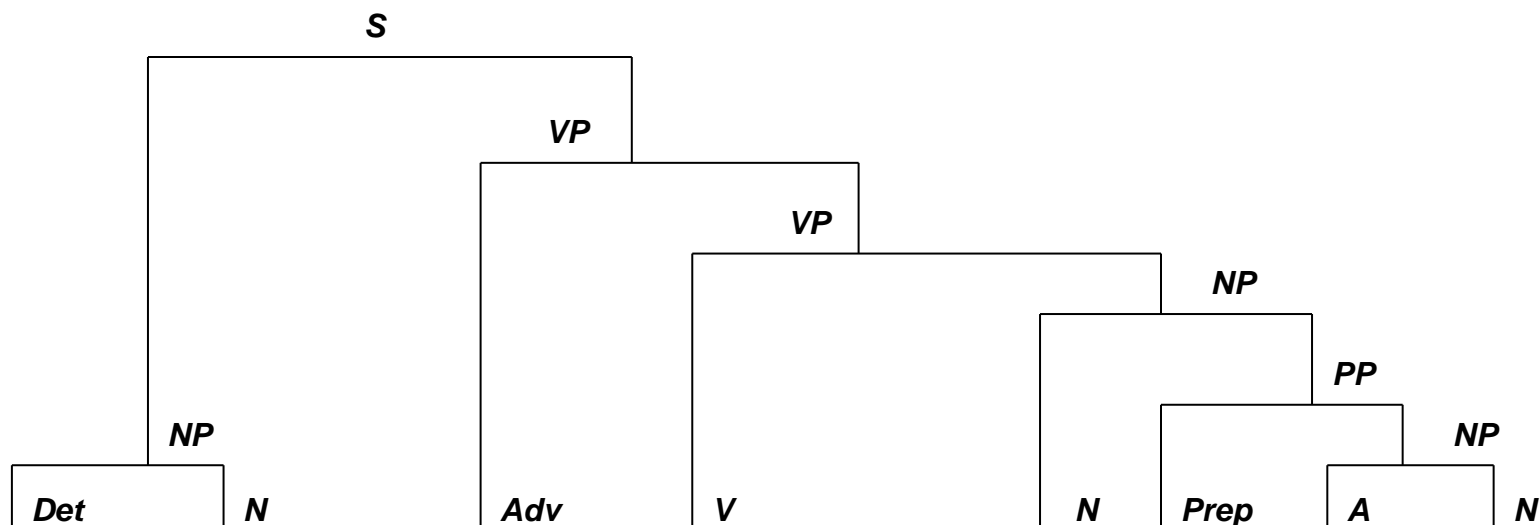


- Для анализа ЕЯ обычно используется  
*размеченная система составляющих:*  
упорядоченная тройка  $\langle C, W, \varphi \rangle$ ,  
где  $C$  – система составляющих,  
 $W$  – множество *меток* различных *грамматических классов*, т.н. *фразовые категории*,  
 $\varphi$  – отображение  $C$  во множество всех непустых подмножеств  $W$ ,  
т.е. список пар «*составляющая + метка/метки, приписанные данной составляющей*»
- Из этой лингвистической модели возникли формальные грамматики Н. Хомского  
(*генеративная лингвистика*)

# РАЗМЕЧЕННОЕ ДЕРЕВО СОСТАВЛЯЮЩИХ



Метки:	<i>S</i> – предложение	<i>Det</i> – детерминатив
типы	<i>N</i> – существительное	<i>V</i> – глагол
	<i>NP</i> – именная группа	<i>VP</i> – глагольная группа
фраз	<i>A</i> – прилагательное	<i>Adv</i> – наречие
	<i>Prep</i> – предлог	<i>PP</i> – предложная группа



**(Эти школьники) (скоро (напишут (диктант (по (русскому языку))))**

Правильные синтаксические структуры фиксируются КС-грамматикой

# ФОРМАЛИЗАЦИЯ: ГРАММАТИКА ПРЕДЛОЖЕНИЯ

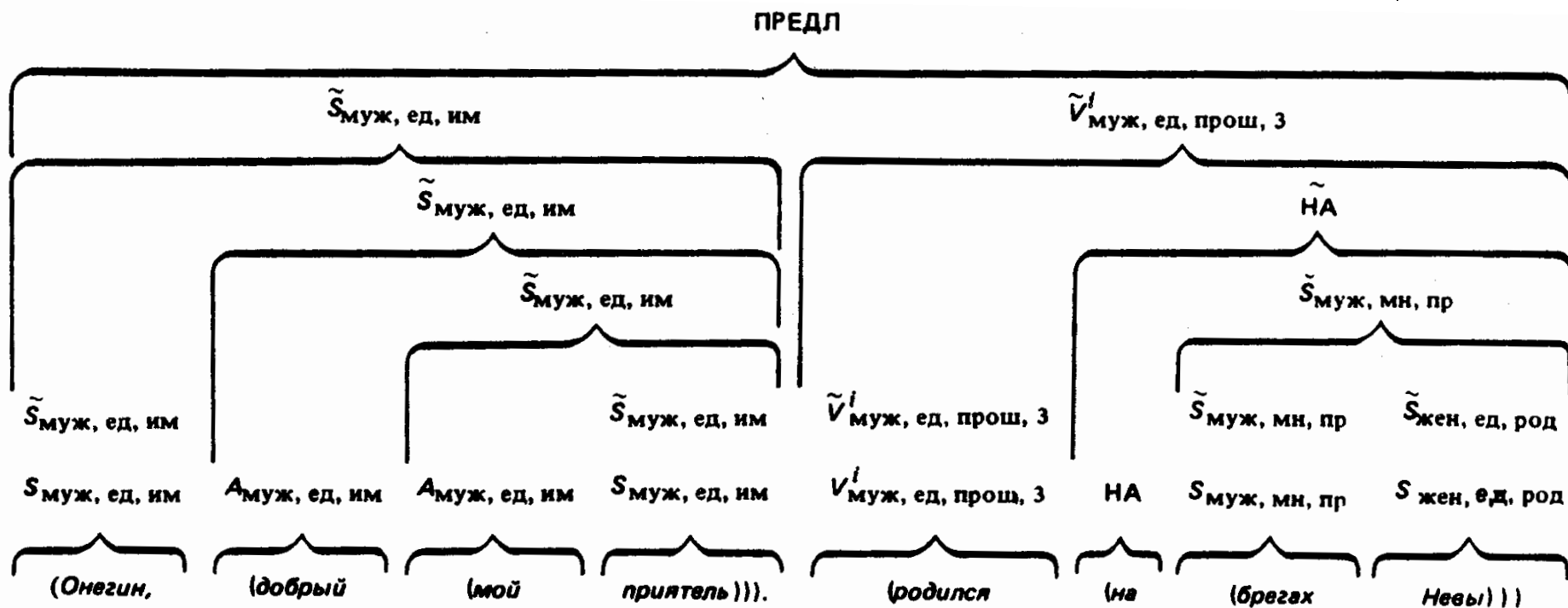


- Формальная *КС-грамматика* (по Н.Хомскому) должна описывать структуру всех грамматически правильных фраз
- КС-грамматика для примера дерева (слайд 12):
  - $S \rightarrow \underline{NP} \underline{VP}$
  - $NP \rightarrow N \mid \underline{A} N \mid \underline{Det} N \mid \underline{N PP}$
  - $VP \rightarrow V \mid \underline{V NP} \mid \underline{Adv VP}$
  - $PP \rightarrow \underline{Prep NP}$

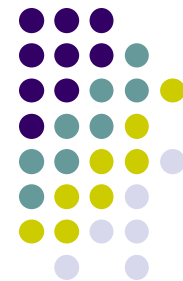
Обозначения: **Sentence**, **Noun**, **NounPhrase**, **Determiner**,  
**Verb**, **VerbPhrase**, **Preposition**, **PrepositionPhrase**,  
**Adjective**, **Adverb**,

- $S, NP, VP, A, N, Det...$  – *Нетерминалы*, соответствуют типам фраз и частям речи;  
*Терминалы ?*

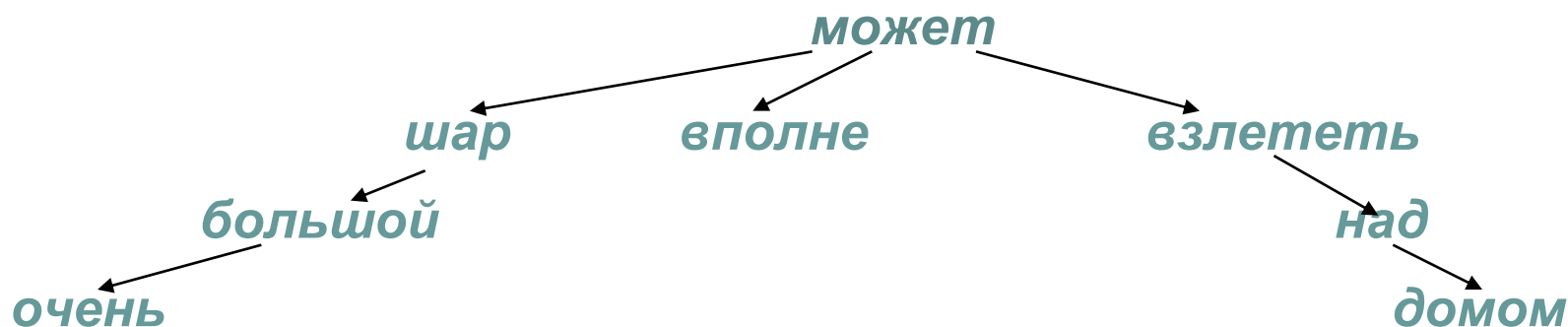
# ДЕРЕВО СОСТАВЛЯЮЩИХ: ДОПОЛНИТЕЛЬНЫЙ ПРИМЕР



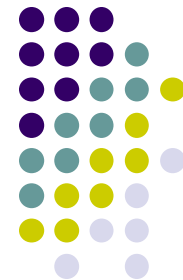
# СТРУКТУРЫ ЗАВИСИМОСТЕЙ: ФОРМАЛИЗАЦИЯ



- Предложение рассматривается как **множество** словоформ  $S = \{w_1, w_2, \dots, w_N\}$
- *Синтаксическая зависимость* – бинарное антисимметричное отношение  $R$  подчинения между его элементами, образующее корневое дерево (неупорядоченное)
  - связность полной структуры предложения
  - нетранзитивное отношение (хотя можно говорить об опосредованном **подчинении**)
- *Дерево зависимостей* (связь подчинения):



# ДЕРЕВЬЯ ЗАВИСИМОСТЕЙ



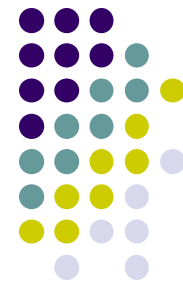
- Дерево зависимостей (подчинения) предложения:
  - ✓ узлы – слова (**корень дерева** – сказуемое)
  - ✓ дуги – подчинительная связь (зависимость)
- **Особенность:** дерево предложения должно быть дополнено информацией о его линейной структуре (т.е. задан порядок слов) – в отличие от деревьев составляющих, которые отражают одновременно синтаксическую и линейную структуру предложения.
- Дерево зависимостей можно изобразить, связав слова на прямой направленными дугами зависимости (подчинения); причем все дуги по одну сторону от прямой:

Очень    большой    шар    вполне    может    взлететь    над    домом

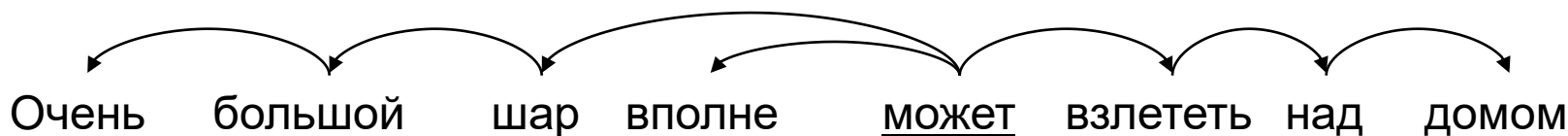
A diagram illustrating dependency arcs for the sentence "Очень большой шар вполне может взлететь над домом". The words are arranged horizontally. Above them, several curved arrows (arcs) represent dependencies: from "Очень" to "большой", from "большой" to "шар", from "шар" to "может", from "может" to "взлететь", from "взлететь" to "над", and from "над" to "домом". All arcs are directed downwards from left to right, staying entirely above the baseline of the words.



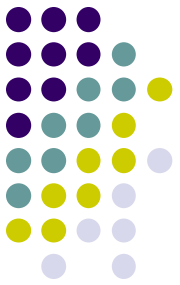
# ДЕРЕВЬЯ ЗАВИСИМОСТЕЙ: ПРОЕКТИВНОСТЬ



- Проективность: некая «правильность» фразы  $S$
- Дерево зависимостей  $\langle S, R \rangle$  для цепочки слов  $S = (w_1, w_2, \dots, w_N)$  называется *проективным*, если для  $\forall \alpha, \beta, \gamma$  – точек цепочки, таких что если  $\alpha \rightarrow \beta$  и  $\gamma$  находится между  $\alpha$  и  $\beta$ , следует, что  $\gamma$  зависит от  $\alpha$ :  $\alpha \rightarrow \gamma$
- *Проективность* означает возможность изобразить зависимости слов  $S$  на прямой так, что одновременно:
  - а) ни одна из дуг не пересекает другую дугу,
  - б) никакая дуга не накрывает вершину (корень дерева)

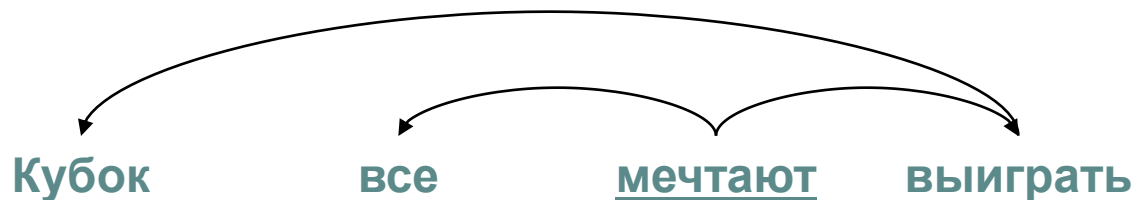


# ДЕРЕВЬЯ ЗАВИСИМОСТЕЙ: СЛАБАЯ ПРОЕКТИВНОСТЬ

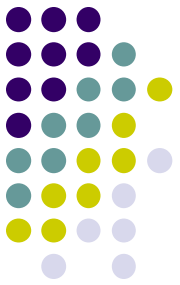


- Содержательный смысл проективности: синтаксически связанные слова близки к друг другу в цепочке слов предложения.
- Большинство правильных предложений русского языка проективны.
- Однако возможен также случай **слабой проективности**.
- Дерево **слабопроективно**, если ни одна из дуг не пересекает другую дугу, но допускается накрывание дугой корневой вершины.

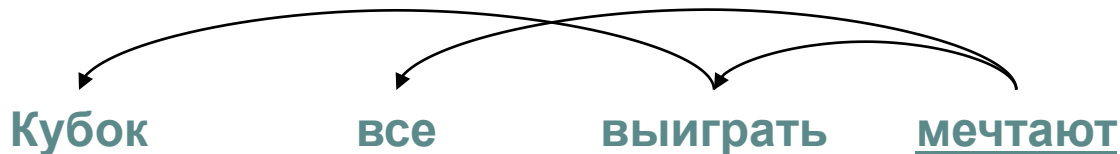
Пример слабой проективности:



# ДЕРЕВЬЯ ЗАВИСИМОСТЕЙ: НЕПРОЕКТИВНОСТЬ



- Непроективные предложения встречаются в художественной литературе, в разговорной речи чаще на языках со свободным (или относительно свободным) порядком слов.
- Такие предложения усложняют синтаксический анализ (и сложнее воспринимаются человеком).
- Примеры:
  - *Я памятник себе воздвиг нерукотворный*
  - *Кубок все выиграть мечтают*



# РАЗМЕЧЕННЫЕ ДЕРЕВЬЯ ЗАВИСИМОСТЕЙ



Для анализа ЕЯ обычно используется

*размеченное дерево зависимостей :*

упорядоченная четверка  $\langle S, R, W, \varphi \rangle$  ,

где  $S$  – множество слов предложения,

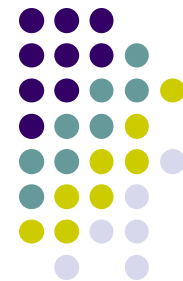
$R$  – отношение, которым задается дерево  
зависимостей для  $S$

$W$  – множество меток возможных  
типов синтаксических связей

$\varphi$  – отображение множества дуг дерева во  
множество  $W$ , т.е.

список пар «дуга дерева + метка типа связи»

# ТИПЫ СИНТАКСИЧЕСКИХ СВЯЗЕЙ



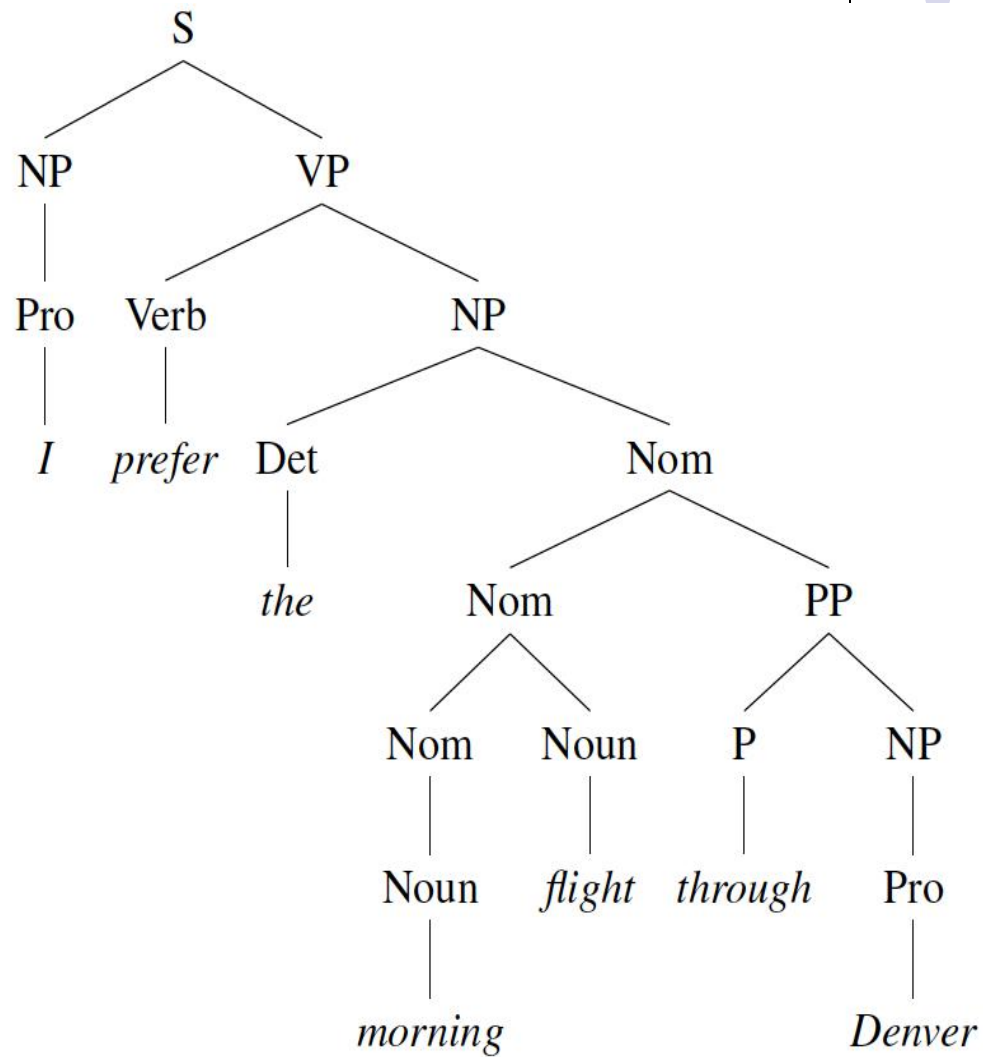
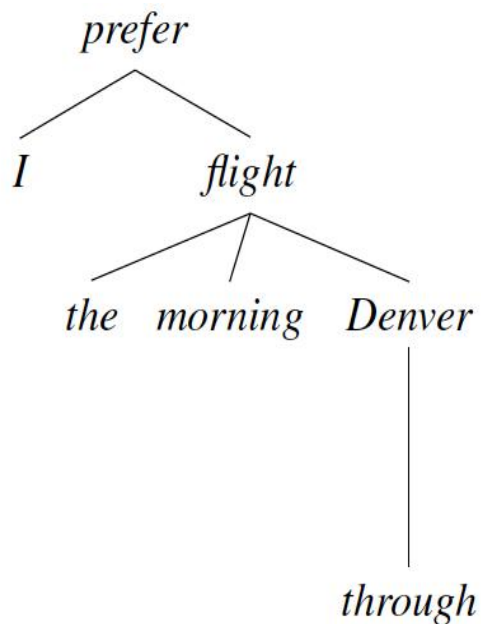
- Количество и набор типов зависит от конкретной модели синтаксического анализа.
- Наиболее распространенные виды:
  - Прямообъектное: *уделить* → *внимание*, *вижу* → *лес*
  - Определительное: *очень* ← *хорошо*, *важный* ← *вопрос*
  - Отпредложное: *в* → *здание*, *хлеб* → *с* → *маслом*
  - Предикат (сказуемое) и субъект (подлежащее):  
*спасатели* ← *обнаружили*
  - Посессивное: *книга* → *врача*
  - Аппозитивное (приложение): *диван* ← *кровать*
  - Количественное: *пять* ← *машин*
  - Обстоятельственное: *лежать* → *на* → *полу*
  - Ограничительное: *не* → *для* → *всех*

# СРАВНЕНИЕ МОДЕЛЕЙ СА: ОБЩИЕ СВОЙСТВА



- Для любого предложения только некоторые (размеченные) деревья составляющих и некоторые (размеченные) деревья зависимостей **правильны** с лингвистической точки зрения  $\Rightarrow$  нужны **грамматики**:
  - грамматики составляющих (например, *КС-грамматики*)
  - грамматики зависимостей
- Для некоторых предложений\фраз ЕЯ возможно более одной правильной синтаксической структуры – это случай **синтаксической омонимии**, которая не может быть разрешена на этапе самого СА (разные смыслы фразы)
  - *Он встретил брата в костюме /в коридоре*
  - *Я видел его молодым...      Мать любит дочь.*(принципиальная **неоднозначность** грамматик ЕЯ)

# СРАВНЕНИЕ ДЕРЕВЬЕВ СА: ПРИМЕР



# ОСОБЕННОСТИ СТРУКТУР СОСТАВЛЯЮЩИХ



## Достоинства:

- Естественное представление неподчинительных отношений: *(картонка и (маленькая собачонка))*
- Фиксируют в явном виде разные словосочетания/фразы

## Недостатки:

- Синтаксическая связь – только вложение фраз
- Неоднозначности членения на фразы:  
*(древние (стены города))* и *((древние стены) города)*
- Не позволяют представлять разорванные синтаксические единицы и непроективные структуры, в частности, вопросит. предложения: *Which book* *did the student* *read?*
- Проблемы представления сложных предложений

Подходят для описания синтаксиса языков со **строгим**  
**(жестким) порядком слов** (английский и др.)



# ОСОБЕННОСТИ ДЕРЕВЬЕВ ЗАВИСИМОСТЕЙ



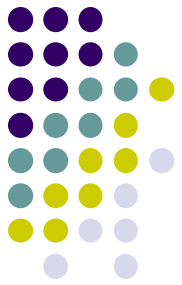
## Достоинства:

- Возможность представления непроективных структур
- Возможность отображать разнотипные синтаксические связи (но только между словами)

## Недостатки:

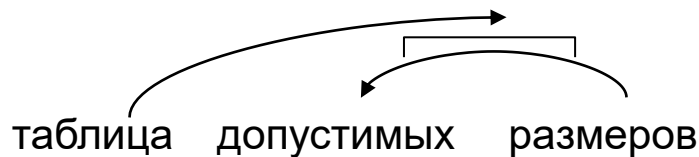
- Неоднозначности в отображении неподчинительных (сочинительных) отношений: *красивый и умный*
- Не позволяют отобразить связи:
  - разноуровневых единиц (крупнее слово), например, конструкции с вспомогат. глаголом: *будет читать*
  - двойного подчинения и приложений: *директор Иванов*

Подходят для языков с достаточно свободным порядком слов (русский, испанский и др.)



# КОМБИНИРОВАННАЯ МОДЕЛЬ СА

- Попытки преодолеть ограничения подходов
- Гладкий А. (1985 г.) – теория **синтаксических групп**
- **Синтаксическая группа**: множество слов (фраза), которое вступает в отношение зависимости целиком, а не посредством одного из входящих в него слов,
- Пример гибкости комбинированной модели:
  - (а) **таблица допустимых размеров**  
(таблица, в которую сведены допустимые размеры)  
синтаксическая группа: **допустимых размеров**
  - (б) **таблица допустимых размеров**  
(таблица, размеры которой допустимы)



# КОМБИНИРОВАННЫЕ СТРУКТУРЫ СОСТАВЛЯЮЩИХ И ЗАВИСИМОСТЕЙ



*Он любил ходить без шапки.*

Синтаксич. группы с внутренней иерархией и без таковой, например: отсутствие внутренней иерархии в предложном сочетании

*Без шапки любил ходить Иван.*

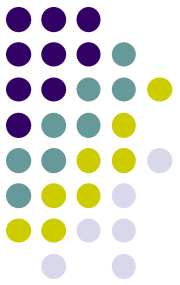
Возможность установления подчинительной связи между группами в целом

*Без шапки Иван ходить не любил*

Допустимость разрывных групп

Примеры А.Гладкого

# ПАРСЕРЫ ДЛЯ СА: ПОДХОДЫ К ПОСТРОЕНИЮ



*Парсер* – синтаксический анализатор

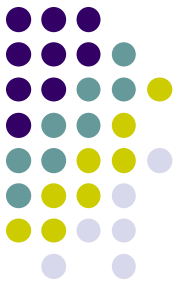
На входе: предложение текста, являющееся результатом морфологического анализа словоформ !)

На выходе: *синтаксическое дерево* предложения

Подходы к построению парсеров:

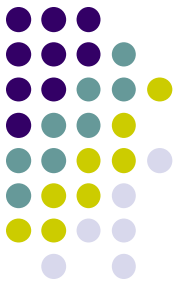
- Подход, базирующийся на правила и словарях  
модель СА **взаимосвязано включает**:
  - способ представления синтаксической структуры
  - способ описания грамматических правил:  
грамматика составляющих / грамматика зависимостей,  
(строится экспертами или автоматизиров. по корпусам)
  - метод/алгоритм синтаксического анализа
- Подход на основе статистики и машинного обучения
- Гибридные подходы

# СТРАТЕГИИ АНАЛИЗА ДЛЯ ГРАММАТИК СОСТАВЛЯЮЩИХ

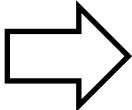


- Алгоритмы СА на основе КС-грамматик:  
Нисходящие ( *top-down* ) или Восходящие ( *bottom-up* )  
В общем случае – недетерминированный разбор  
с возвратами, экспоненциальная сложность
- Алгоритм Кока-Янгера-Касами ( *СΥΚ* )
  - КС-грамматики в *нормальной* форме Хомского:  
правила вида  $A \rightarrow BC$  и  $A \rightarrow \gamma$
  - Разбор снизу-вверх
  - Полиномиальная сложность:  $O(|G| \times n \times n)$ ,  
 $n$  – длина предложения,  $|G|$  - мощность грамматики
- Алгоритм Эрли
  - Не накладывает ограничений на грамматику
  - Разбор сверху-вниз
  - Кубическая сложность
- Проблема: неоднозначность грамматик, омонимия

# РАЗРЕШЕНИЕ СИНТАКСИЧЕСКОЙ ОМОНИМИИ ДЛЯ КС-ГРАММАТИК



- Применение статистики синтаксических разборов
- На основе корпуса подсчитать вероятность каждого правила грамматики:

$P = \{...$		$P = \{...$
$NP \rightarrow DT\ NN$		$NP \rightarrow DT\ NN$ 0,3
$NP \rightarrow DT\ ADJ\ NN$		$NP \rightarrow DT\ ADJ\ NN$ 0,6
$NP \rightarrow NN\ NN$		$NP \rightarrow NN\ NN$ 0,1
$... \}$		$... \}$

- В итоге: КС- грамматика с вероятностями
- Вероятность дерева разбора определяется перемножением вероятностей правил, примененных при его построении
- Выбирается наиболее вероятное дерево

# СТРАТЕГИИ АНАЛИЗА ДЛЯ ГРАММАТИК ЗАВИСИМОСТЕЙ



- Рассматриваются деревья зависимости, для которых выполняется свойство проективности
- Перебор на основе графов: среди всех возможных деревьев ищется правильное
- Метод на основе переходов: анализ предложения слева направо, поиск связей между соседними словами
- Метод фильтров:
  - Порождаются всевозможные синтаксические связи слов
  - Отбрасываются ошибочные и избыточные связи путем применения фильтров (например, правил согласования)
  - Фильтры: условия правильно построенных деревьев
- На практике для эффективности часто применяется:  
*синтаксическая сегментация (Syntactic chunking)*  
путем установления **высоковоероятных локальных связей**

# СИНТАКСИЧЕСКАЯ СЕГМЕНТАЦИЯ



Может предшествовать построению синт. дерева,  
быть начальным этапом СА – *Syntactic chunking*,  
*частичный синтаксический анализ*, выделение:

- простых предложений в составе сложных для проведения их независимого синтаксического анализа.
- локальных синтаксич. групп – именных, глагольных и др. словосочетаний на базе *высоковоероятных локальных связей*.

Примеры:

- {Студент московского университета} прочитал {весьма интересную английскую статью} и {скоро будет делать} {краткий доклад}.
- Артур указал на {эту девочку}, {предложившую принести} {чистые тарелки и ложки}.
- [Олег узнал его], [он {видел этого человека} {два года} назад].

Вычислительная сложность –  $O(n)$



# ПРАВИЛА УСТАНОВЛЕНИЯ ЛОКАЛЬНЫХ СВЯЗЕЙ



Пример: Установление высоковероятной локальной связи прилагательное+существительное –  $A \ N$   
*приветливый взор, открытый взору*

- Если у  $N$  и  $A$  совпадают род, число, падеж (согласование), то сделать существительное  $N$  главным и установить связь  $A \leftarrow N$
- Если  $N$  является неизменяемым, то сделать его главным и установить связь  $A \leftarrow N$
- Если прилагательное  $A$  является отглагольным, то сделать его главным и установить связь  $A \rightarrow N$
- Если прилагательное является неизменяемым, то сделать  $N$  главным словом и установить связь  $A \leftarrow N$

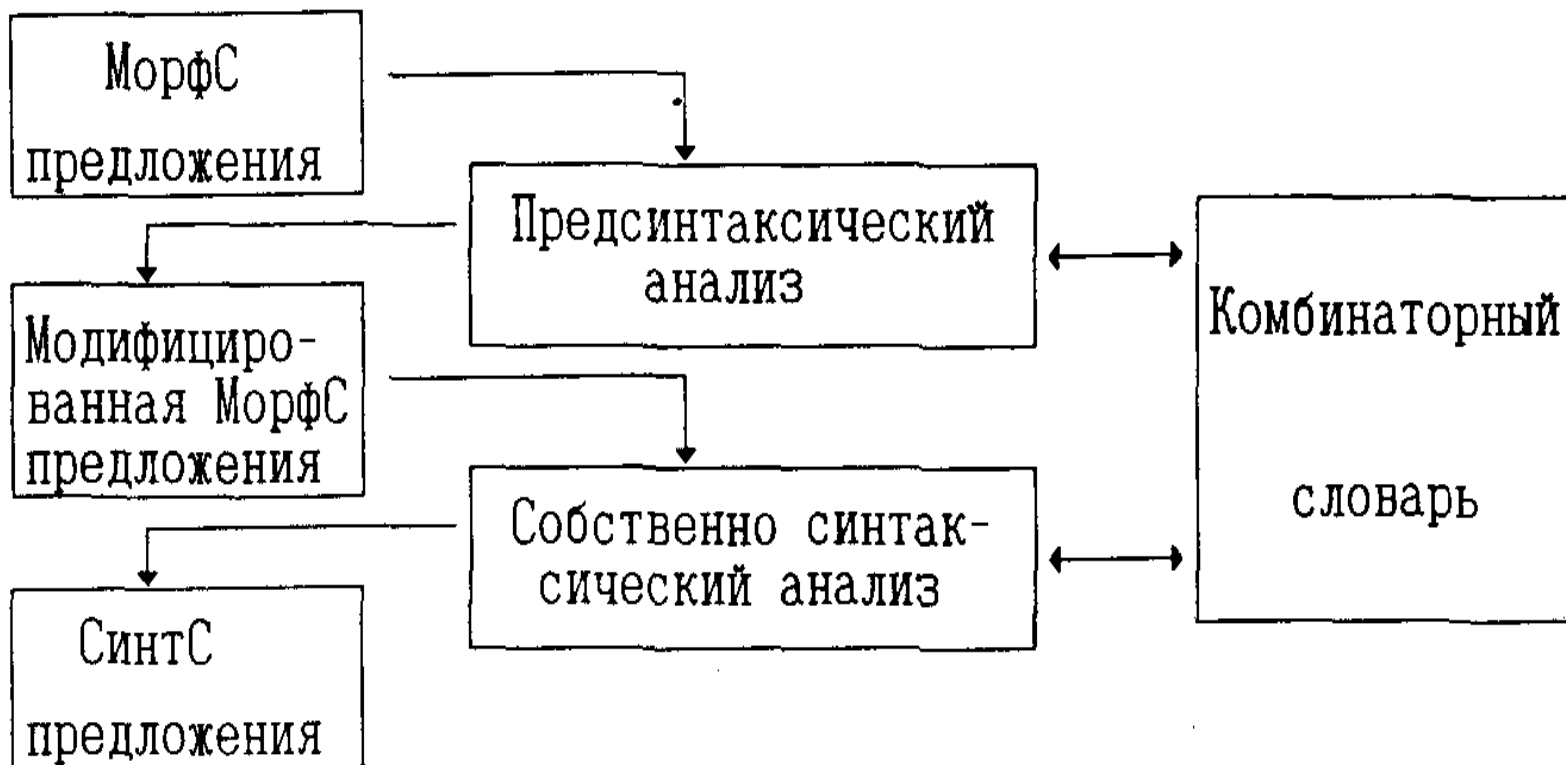
Возможны другие высоковероятные локальные связи

# ПАРСЕРЫ НА ПРАВИЛАХ: СИСТЕМА ЭТАП

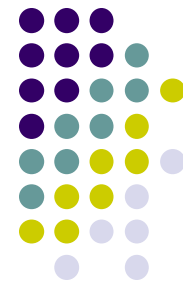


- Система *Этап* разрабатывалась для машинного перевода с 1970 гг., одна из первых значимых систем в настоящее время – *Этап-3* <http://proling.iitp.ru/etap/>
- Несколько ЕЯ: русский, французский, английский
- База: Лингвистическая теория (модель) «Смысл $\Leftrightarrow$ Текст»  
поверхностный и глубинный синтаксис
- Парсер системы: деревья и грамматика зависимостей, много лингв. правил анализа (декларат. представление)
- Синтаксическая информация представлена также в ТКС – *толково-комбинаторном* словаре:  
модели управления слов-предикатов,  
т.е. описание их синтаксических валентностей, в частности: падеж актантов  
(*актант* – заполнитель валентности: слово/фраза)

# СИСТЕМА ЭТАП: СИНТАКСИЧЕСКИЙ АНАЛИЗ

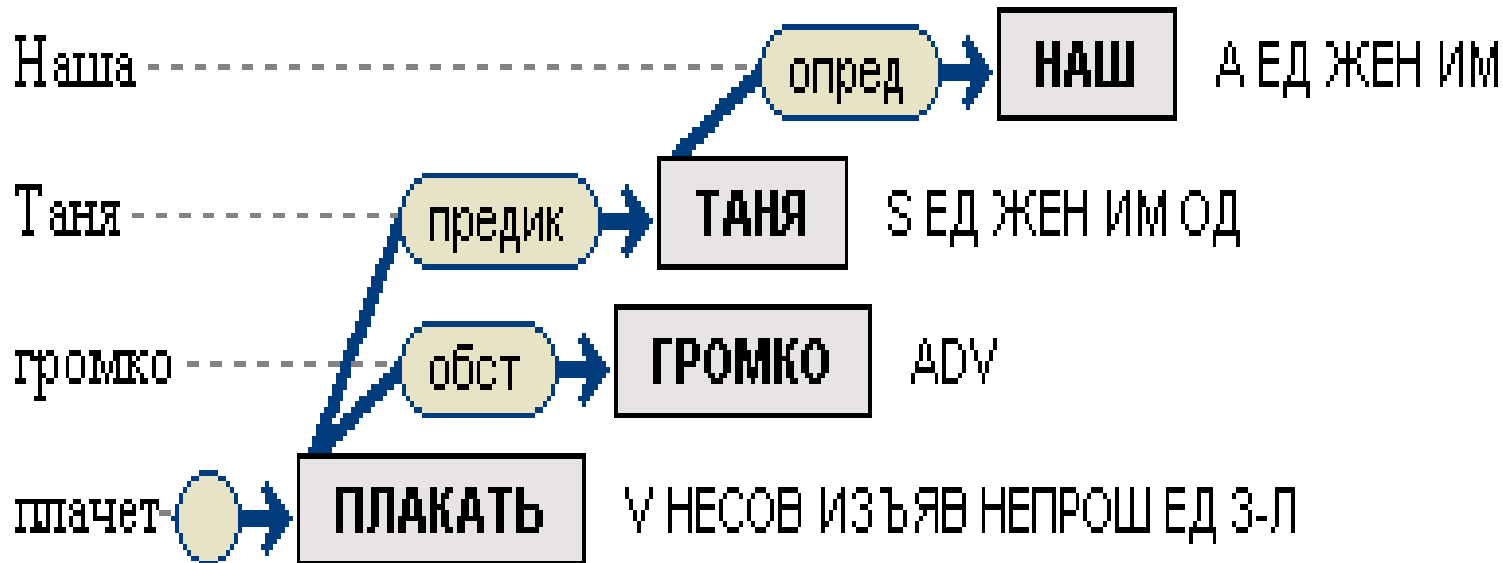


# ЭТАП: ПРИМЕР СИНТАКСИЧЕСКОЙ СТРУКТУРЫ



Размеченное *дерево* зависимостей:

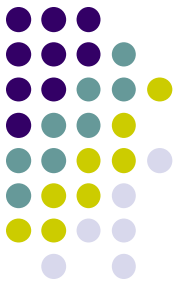
- узлы дерева – слова предложения
- каждая дуга дерева помечена именем синтаксического отношения : *СинтО* (поверхностный синтаксис)



# ЭТАП: СХЕМА АНАЛИЗА



# ПАРСЕРЫ НА ПРАВИЛАХ ДЛЯ РЯ: *ДИАЛИНГ-АОТ*



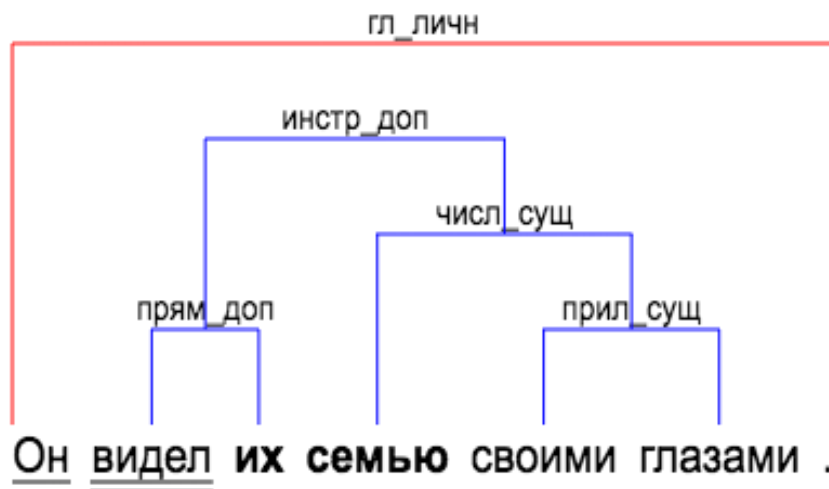
Проект *Диалинг-АОТ* ([aot.ru](http://aot.ru)) лингвистич. анализа русскоязычных текстов (1998-2001 гг.), открытый код, инженерный подход (на правилах), веб-интерфейс

- Модуль графематики: токенизация и сегментация на предложения, свертка словарных словосочетаний и др.
- Морфологический модуль: для каждой словоформы – множество морфологических омонимов
- Постморфологический анализ на правилах
- Модуль синтаксического анализа **SynAn**
  - базируется на понятии *синтаксической группы* – комбинированная, гибридная модель синтаксиса
  - синтаксические правила представлены процедурно
  - Используются **модели управления** слов-предикатов  
*Ударить – Кто? Кого? Чем?*

# АОТ: ОСОБЕННОСТИ СА



- Не ставится цель получить полную синтаксическую структуру предложения, а только формирование различных синтаксических групп слов
- На начальном этапе – *фрагментационный анализ* (= синтаксической сегментации): деление предложения на неразрывные синтаксические единства и установление частичной иерархии:
  - ❖ главные и придаточные предложения (простые)
  - ❖ причастные и деепричастные обороты



Что здесь неверно?

# АОТ: СИНТАКСИЧЕСКИЕ ГРУППЫ (39 типов)



Тип	Название	Пример
Количественная группа (последовательность числительных)	<b>КОЛИЧ</b>	двадцать восемь
Последовательность чисел	<b>СЛОЖ-ЧИСЛ</b>	12,3, II-III
Группа существительного, пре- модифицированная одним или несколькими прилагательными	<b>ПРИЛ-СУЩ</b>	длинная тяжелая дорога, идуший человек
Группа существительного, премодифицированная наречным числительным	<b>НАР-ЧИСЛ- СУЩ</b>	много ребят, мало стульев
Группа существительного, премодифицированная числительным	<b>СУЩ-ЧИСЛ</b>	восемь попугаев, два человека
Предложная группа	<b>ПГ</b>	в дом, на холме
...		



# ПАРСЕРЫ НА ПРАВИЛАХ ДЛЯ РЯ:

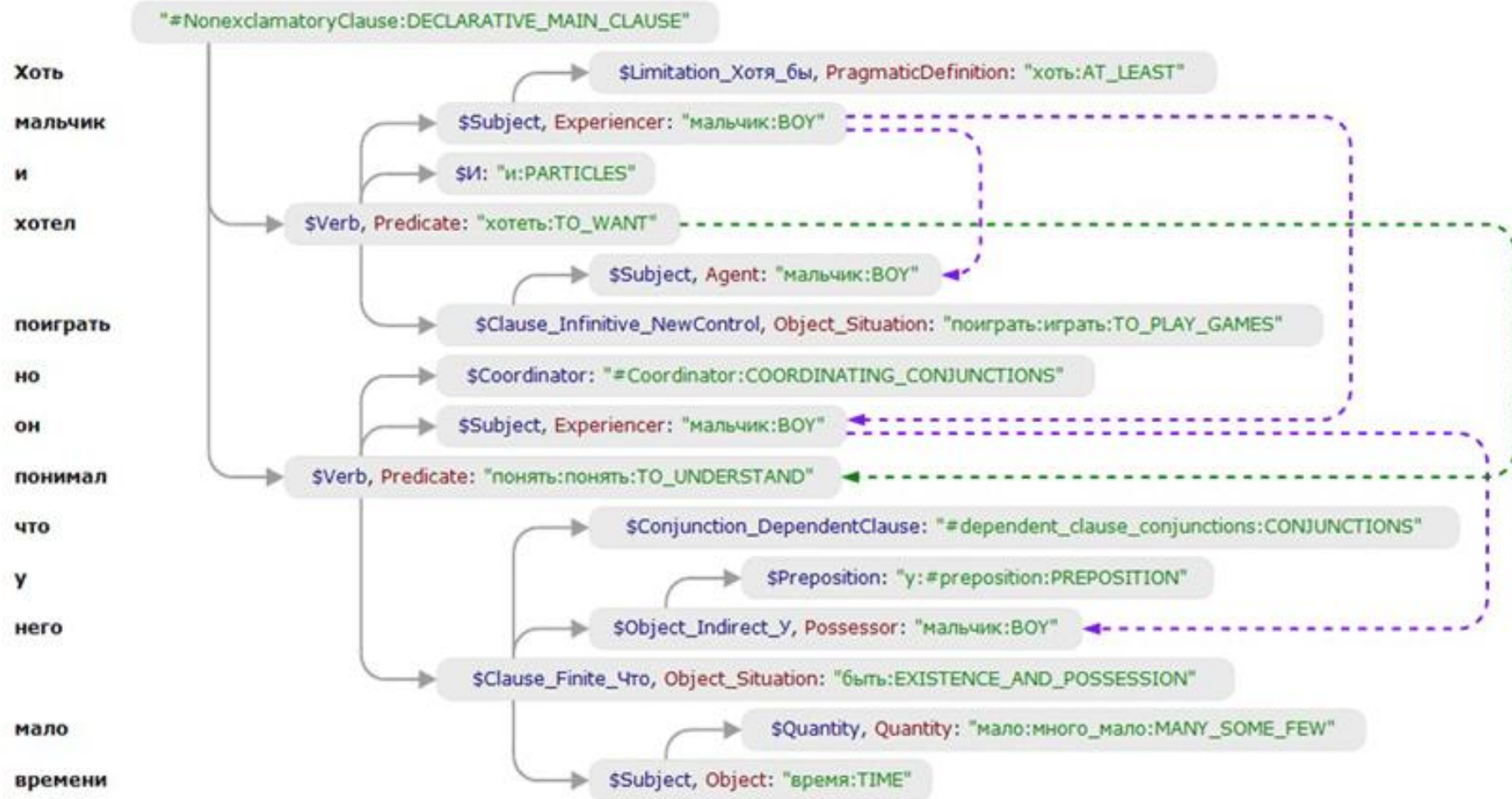
## *Compreno*



- Разрабатывается в АВВУУ более 30 лет
- Объемлющая **система машинного перевода**, построенная на основе перевода любого человеческого языка на универсальный язык понятий и обратно.
- Включает в себя все основные этапы обработки текстов: морфологический, синтаксический и семантический.
- Синтаксический анализ на основе **грамматик зависимостей**, предусматривающих даже непроективные связи.
- Требует громадного объема памяти, не переносим

# Compreno:

## ПРИМЕР СИНТАКСИЧ. АНАЛИЗА

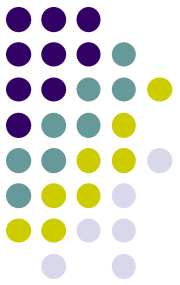


# ПАРСЕРЫ НА ОСНОВЕ ОБУЧЕНИЯ



- Для машинного обучения на основе статистики и необходим корпус с синтаксической разметкой
- Корпуса текстов с синтаксич. разметкой (*Treebank*):
  - для английского языка – *Pen Treebank*
  - для чешского языка – *Prague Dependency Treebank*
  - для русского – *SynTagRus* (синт. подкорпус *НКРЯ*)
- Проект *Universal Dependencies*:
  - более 100 корпусов для 60 языков
- Современные парсеры в открытом доступе, обученные для нескольких языков:
  - *Stanford Parser*
  - *MaltParser*
  - *SyntaxNet*
  - *UDPipe*

# ПРИМЕР СИНТАКСИЧЕСКОЙ РАЗМЕТКИ



- Национальный корпус русского языка (*НКРЯ*):  
*ruscorpora.ru*
- Подкорпус с синтаксической разметкой: *SynTagRus*
  - Разметка корпуса производилась в полуавтоматическом режиме:
  - Обработка предложения морфологическим и синтаксическим анализатором *ЭТАП*
  - Коррекция лингвистом
  - В результате, для каждого предложения:  
правильная морфологическая разметка  
+ единственное, размеченное *дерево зависимостей*
- Важно: качество разметки, подкорпус пополняется

19\_108.pdf (объект «application/pdf») - Mozilla Firefox

Файл Правка Вид Журнал Закладки Инструменты Справка

louk\_nat@... "Ингейт" за... ЕТАР-3 МТ... Яндекс.Но... http://www... russia - По... Переводчи... Поиск в ко... Националь... 19\_108.pdf ...

ruscorpora.ru/syntax/2006/19\_108.pdf

Save a Copy Print Email Search Select Text 170% Download New Reader Now

Bookmarks Signatures Layers Pages

Непременное .....  
условие .....  
существования .....  
открытого .....  
гражданского .....  
общества .....  
свобода .....  
слова .....  
и .....  
прочно .....  
усвоенное .....  
уважение .....  
к .....  
чуждому .....  
мнению .....

опред → НЕПРЕМЕННОЙ [А ЕД СРЕД ИМ]  
предик → УСЛОВИЕ [S ЕД СРЕД ИМ НЕОД]  
1-компл → СУЩЕСТВОВАНИЕ [S ЕД СРЕД РОД НЕОД]  
опред → ОТКРЫТЫЙ [А ЕД СРЕД РОД]  
опред → ГРАЖДАНСКИЙ [А ЕД СРЕД РОД]  
квазиагент → ОБЩЕСТВО [S ЕД СРЕД РОД НЕОД]  
1-компл → СВОБОДА [S ЕД ЖЕН ИМ НЕОД]  
1-компл → СЛОВО [S ЕД СРЕД РОД НЕОД]  
сочин → И [CONJ]  
обст → ПРОЧНО [ADV]  
опред → УСВАИВАТЬ [V СОВ СТРАД ПРИЧ ПРОШ ЕД СРЕД ИМ]  
соч-союзн → УВАЖЕНИЕ [S ЕД СРЕД ИМ НЕОД]  
1-компл → К [PR]  
опред → ЧУЖОЙ [А ЕД СРЕД ДАТ]  
предл → МНЕНИЕ [S ЕД СРЕД ДАТ НЕОД]

19,877 x 8,889 cm 1 of 1

# ОЦЕНКИ КАЧЕСТВА СИНТАКСИЧЕСКОГО АНАЛИЗА



- Для деревьев составляющих оценивают полноту, точность, F-меру и т.д. верно размеченных в рамках предложения составляющих (правильно указано начало, конец и нетерминальный символ)
- Для деревьев зависимостей оценивают:
  - процент слов, правильно определенных корнем дерева – метрика *unlabeled attachment score (UAS)*
  - процент слов с правильно определенной родительской вершиной и типом зависимости – метрика *labeled attachment score (LAS)*
  - *accuracy* – доля верных ответов
- НО ! На эти оценки влияет качество предшествующих этапов: токенизации, морфологического анализа
- *CoNLL Shared Task*, обучение на *SynTagRus* : 86-92% *LAS*

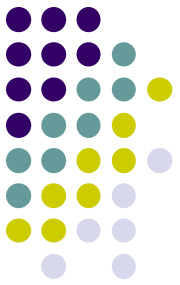
# Stanford Parser : ВЕРОЯТНОСТНЫЕ МОДЕЛИ



<http://nlp.stanford.edu/software/lex-parser.html>

- Несколько версий, изначально – для английского языка, сейчас – для ряда языков: немецкого, итальянского, португальского, болгарского, арабского, китайского и др.
- Версия 3.4 (2014 г.) : *Shift-reduce constituency parser*
  - деревья составляющих, нисходящий анализ
  - КС-грамматика *PCFG*, лексикализация, применение вероятностей и контекстной информации
  - *алгоритм A\** при построении дерева
  - вычислительная сложность  $O(n^5)$  от длины входа
  - средства преобразования в дерево зависимостей
- Нейросетевая версия 3.5: *Neural-network dependency parser*
  - деревья зависимостей + типы синт. связей
  - только для английского и китайского языков

# *MaltParser* : ОБУЧЕНИЕ ДЛЯ СТРУКТУР ЗАВИСИМОСТЕЙ

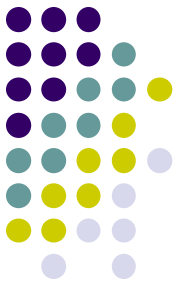


<http://maltparser.org/>

- Первый значимый на основе машинного обучения, для деревьев зависимостей, 2 режима работы:
  - Обучение синтаксической модели: построение правил, предсказывающих действия анализатора
  - Синтаксический анализ на основе обученной модели
- Обученные модели для разных ЕЯ, неплохие результаты даже при небольшом обучающем корпусе
- Первая модель 2011 г. для русского языка:
  - корпус *SynTagRus* (41186 предл., 719957 токенов), 70% для обучения, 15% для настройки параметров
  - предобработка текста: *TreeTagger*, *CSTLemma*
  - тестирование (15% корпуса): *UAS* – 88.0 %  
*LAS* – 82.2 %

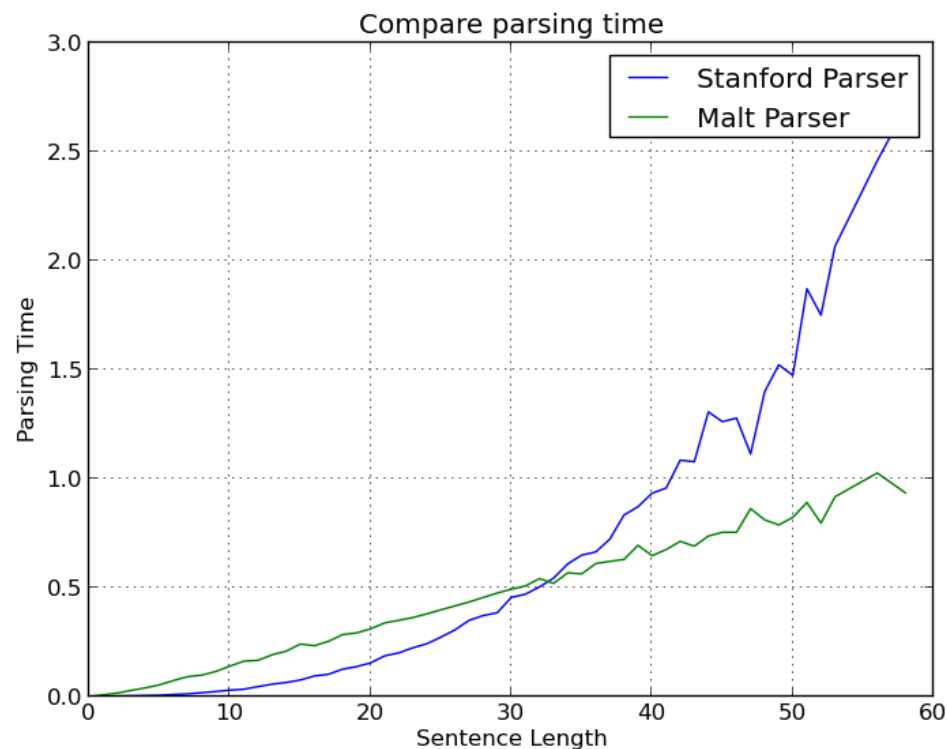
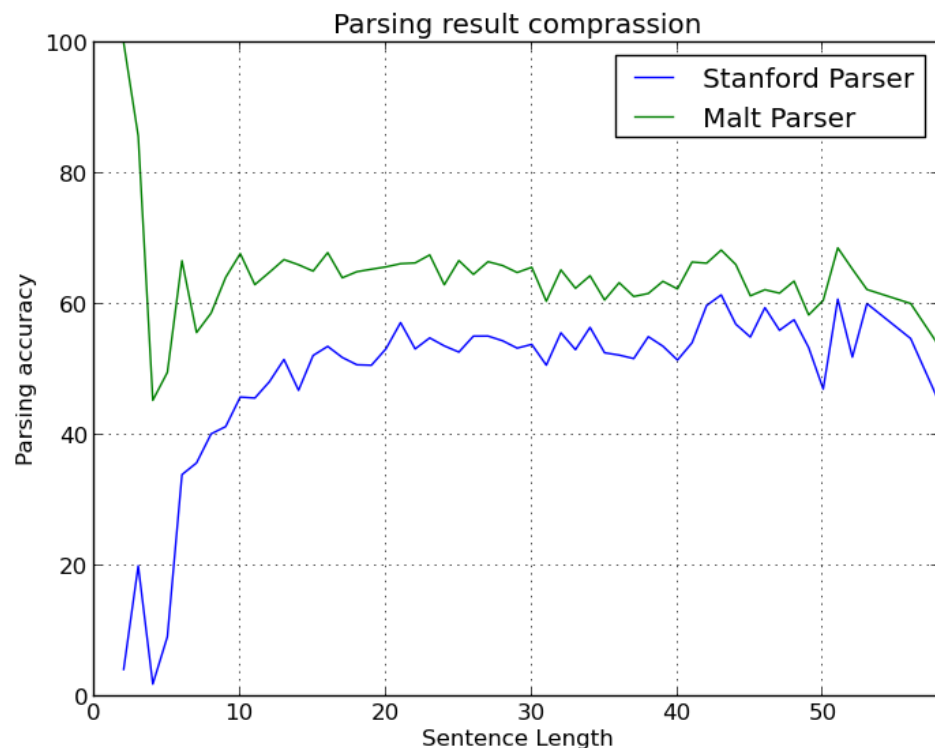


# *MaltParser: АНАЛИЗ*



- ❖ Особенность размеченного дерева зависимостей:
  - узлы соответствуют элементам предложения: словам и знакам препинания
  - специальная (пустая) корневая вершина
- ❖ Детерминированный алгоритм построения дерева, архитектура *transition-based parser*
- ❖ Разбор предложения происходит слева направо, один проход по предложению, используются два стека: для обработанной и необработанной частей предложения
- ❖ Процесс анализа – последовательные **переходы** (4 вида) от одной конфигурации анализа к другой
- ❖ **Предсказатель** (машинный классификатор) определяет переход по текущей конфигурации
- ❖ Вычислительная сложность  **$O(n)$** , по все же относительно долгое время работы парсера

# СРАВНЕНИЕ *Stanford Parser* и *MaltParser* НА ОДНОЙ КОЛЛЕКЦИИ



# *MaltParser* : ОБУЧЕНИЕ

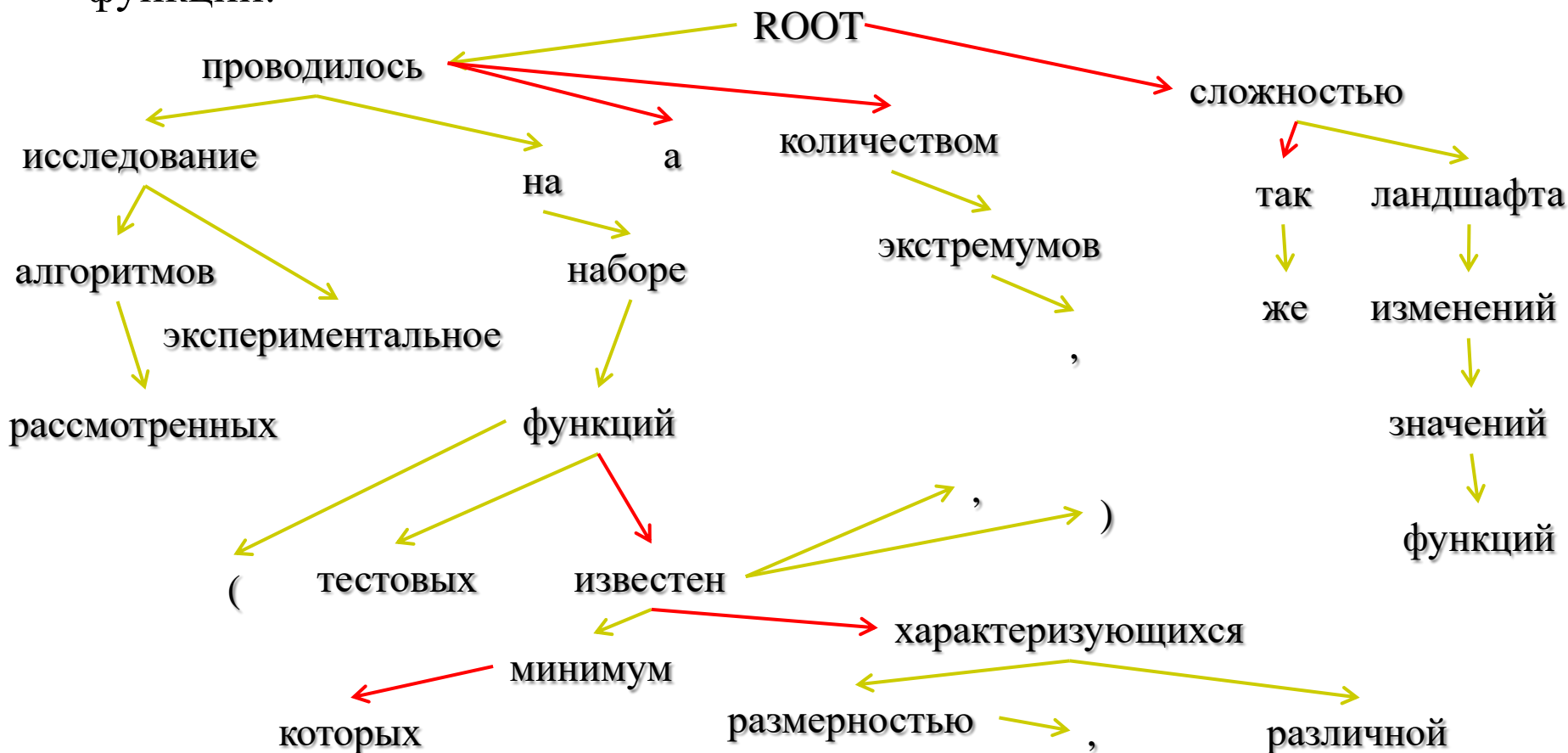


- Построение правил, предсказывающих действия анализатора по *признакам*  
*предыстории разбора*
- Предсказатель фактически опирается на признаки предыстории разбора, соответствующие некоторой частично построенной древесной структуре
- Признаки: части речи, типы зависимостей, леммы некоторых элементов используемых стеков (берутся из синтаксически размеченного корпуса)
- Разбиение предысторий на классы эквивалентности, в зависимости от учитываемых признаков
- Обучение сводится к задаче классификации, методы:
  - *SVM* – метод опорных векторов
  - *MBL* – обучение, основанное на предыстории с использованием  $k$  ближайших соседей

# MaltParser: ПРИМЕР РАЗБОРА



Экспериментальное исследование рассмотренных алгоритмов проводилось на наборе тестовых функций (минимум которых известен), характеризующихся различной размерностью, количеством экстремумов, а также сложностью ландшафта изменений значений функций.



# ПАРСЕР SYNTAXNET

<https://opensource.google.com/projects/syntaxnet>



- Фреймворк для разработки систем анализа ЕЯ, от *Google*, 2 режима работы: Обучение + Синт.анализ
- Машинное обучение модели на основе нейронных сетей, с применением библиотеки *TensorFlow*
- Обученные модели для 40 языков, для построения синтаксических деревьев зависимостей, в них встроены: сегментация на токены, предложения + морфол.анализ
- При обучении использованы корпуса с *UD* разметкой: *Universal Dependencies* – универсальная для всех ЕЯ разметка морфологических признаков и синтаксических зависимостей <https://universaldependencies.org>
- Парсер для РЯ: на корпусе *SynTagRus* качество анализа не сильно превосходит *MaltParser*
- Сравнительно медленная работа, сложность запуска

# *UDPipe* : КОНВЕЙЕР ДЛЯ СИНТАКСИЧЕСКОГО АНАЛИЗА



<http://ufal.mff.cuni.cz/udpipe>

- Развивающийся проект, открытый код
- Обучаемый конвейер для проведения СА, включающий модули токенизации, сегментации на предложения, морфологического и синтаксического анализа
- Построение синтаксических деревьев зависимостей: *transition-based parser* , сложность близка к линейной
- *UDPipe 2.0* : 60 параметров обучения
- Для обучения: корпуса с универсальной для всех ЕЯ разметкой *Universal Dependencies*
- Обученные модели (парсеры) для более 30 языков, 3 модели для РЯ (есть модель для *SynTagRus*)
- Вывод в виде текстового файла с результирующим деревом зависимостей

# UDPipe : ПРИМЕР ВЫВОДА

*Очень большой шар вполне может взлететь  
над домом.*

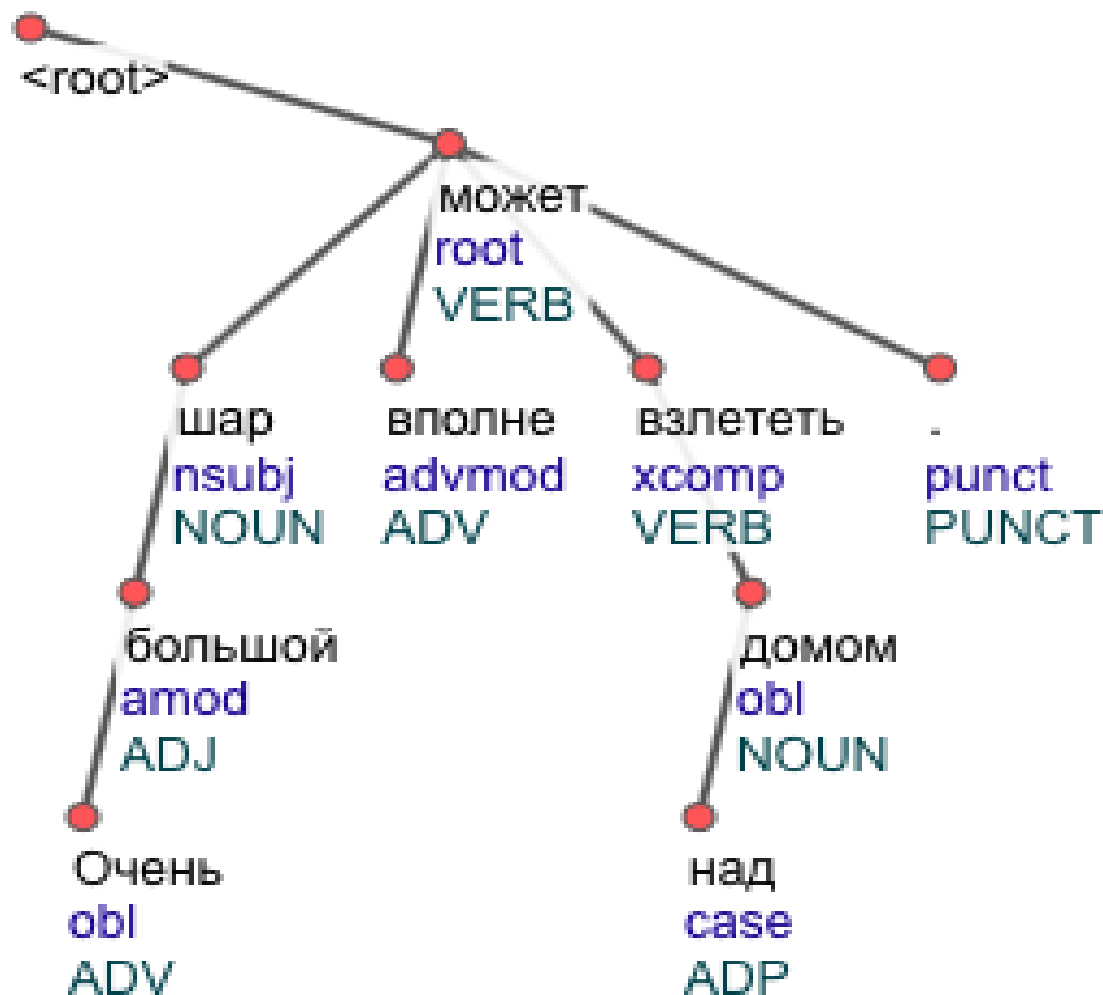


```
1  Очень  очень  ADV  _  Degree=Pos  2  obl  _  _
2  большой  большой  ADJ  _  Case=Nom|Degree=Pos|Gender=
   Masc|Number=Sing  3  amod  _  _
3  шар  шар  NOUN  _  Animacy=Inan|Case=Nom|Gender=Masc|
   Number=Sing  5  nsubj  _  _
4  вполне  вполне  ADV  _  Degree=Pos  5  advmod  _  _
5  может  мочь  VERB  _  Aspect=Imp|Mood=Ind|Number=Sing|
   Person=3|Tense=Pres|VerbForm=Fin|Voice=Act  0  root  _
6  взлететь  взлететь  VERB  _  Aspect=Perf|VerbForm=Inf|Voice
   =Act  5  xcomp  _  _
7  над  над  ADP  _  _  8  case  _  _
8  домом  дом  NOUN  _  Animacy=Inan|Case=Ins|Gender=Masc
   |Number=Sing  6  obl  _  SpaceAfter=No
9  .  .  PUNCT  _  _  5  punct  _  SpacesAfter=\n
```

# UDPipe: ПРИМЕР РАЗБОРА - 1



Очень большой шар вполне может взлететь над домом . }

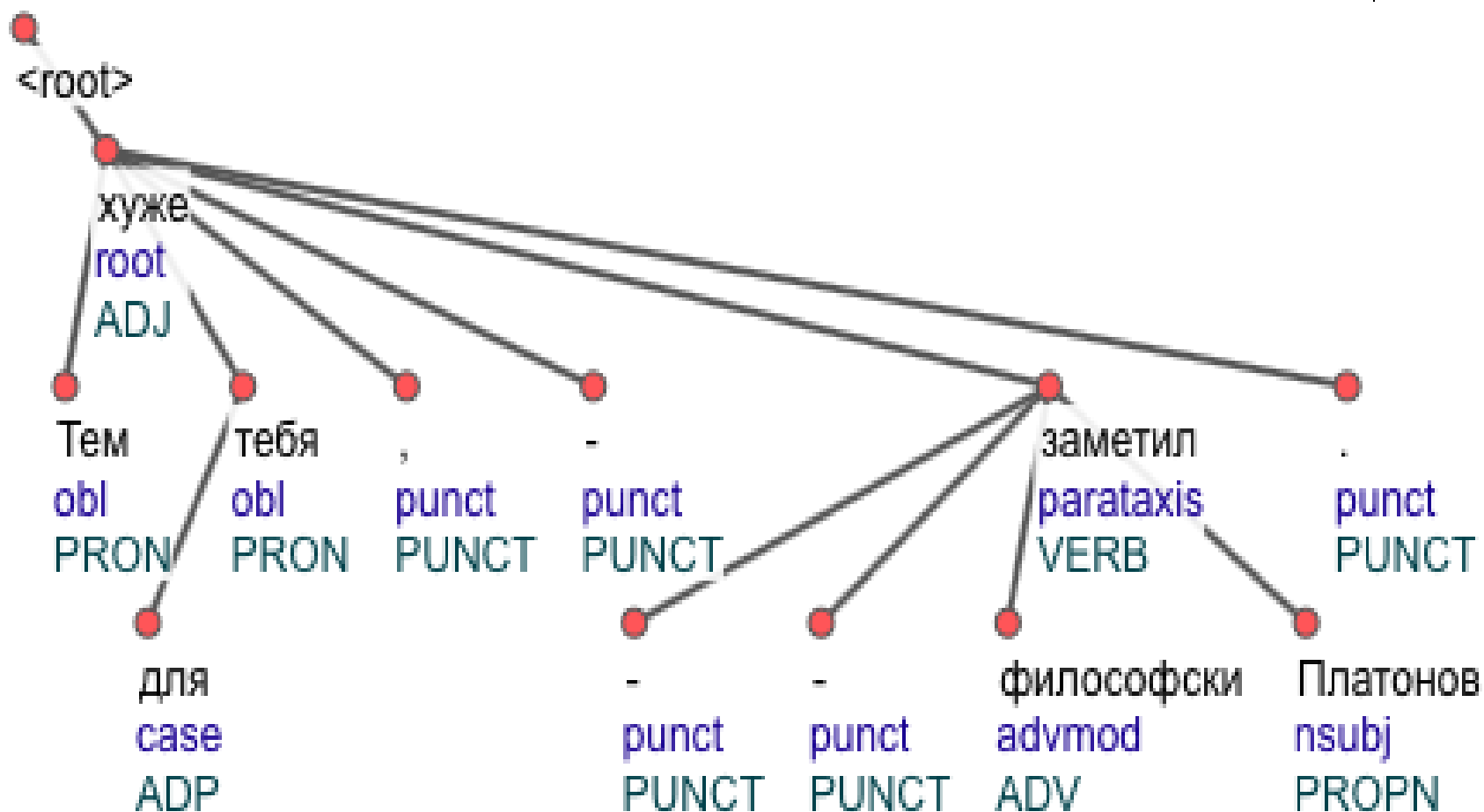




# UDPipe: ПРИМЕР РАЗБОРА - 2



*Тем хуже для тебя, - философски заметил Платонов*



# БИБЛИОТЕКА *NLTK* : ПАРСЕРЫ



- Более 40 синтаксических анализаторов в классе *parse* (<https://www.nltk.org/api/nltk.parse.html>) почти все для английского языка, показывают работу по шагам, есть графический интерфейс (*Demo.py*)
- Для деревьев составляющих:
  - КС-грамматики: разные направления
  - КС-грамматики с вероятностями
- Для деревьев зависимостей, архитектуры:
  - на основе переходов (*transition-based*) : классический алгоритм, а также *MaltParser*
  - на основе графов: разные способы оценки дуг в процессе выделения дерева
- ❖ Генератор предложений по заданной КС-грамматике

# КАЧЕСТВО ПАРСЕРОВ ДЛЯ РЯ



Все парсеры обучены на корпусе *SynTagRus*

Оценивается точность предварительной токенизации и сегментации на предложения

Оценка качества синтаксического анализа:

- правильность корня предложения (*UAS*)
- точность хозяина (предка), от которого есть зависимость
- точность типов всех синтаксических связей

(абсолютно для всех предложений и относительно только верно выделенных при сегментации и токенизации)

парсер	предл.	токены	корень		хозяин		все связи	
			абс.	отн.	абс.	отн.	абс.	отн.
MaltParser	95.27%	77.06%	51.63%	66.99%	26.04%	33.79%	2.64%	3.43%
SyntaxNet	96.14%	84.72%	77.71%	<b>91.73%</b>	64.31%	74.64%	16.20%	18.70%
UDPipe	<b>99.54%</b>	<b>96.24%</b>	<b>87.32%</b>	90.73%	<b>81.17%</b>	<b>84.34%</b>	40.70%	42.29%

# ЗАКЛЮЧЕНИЕ



- Этап синтаксического анализа текста – реально сложный этап, и актуальной остается задача улучшения качества работы парсеров
- Точность разбора зависит от нескольких факторов:
  - качества предшествующих токенизации, сегментации и морфологического анализа
  - использования или нет пунктуации
- Кроме точности СА важна скорость работы парсера
- Существенная проблема – синтаксическая омонимия, чем длиннее предложение, тем в среднем больше вариантов разбора, а выявление из них верного зависит от семантики текста

## СПАСИБО ЗА ВНИМАНИЕ

# РЕЗУЛЬТАТЫ ПАРСЕРА AOT ?

<http://aot.ru/demo/synt.html>



- Эти школьники скоро напишут диктант по русскому языку
- Очень большой шар вполне может взлететь над домом
- Студенты факультета читают рекомендованную литературу по дискретной математике
- Кубок все мечтают выиграть
- Кубок все выиграть мечтают
- Мы увидели больного врача Сидорова
- Он встретил брата в костюме /в коридоре
- Глокая куздра штеко будланула бокра и курдячит бокрёнка

# ДОМАШНЕЕ ЗАДАНИЕ № 3



На выбор варианты:

- A. Исследование различных типов коллокаций на базе ресурса *RuWac*
- B. Сравнительный анализ возможностей двух парсеров для русского/английского языка
- C. Составление программы для синтаксической сегментации текстов на РЯ (на основе правил выявления высоковероятных синтаксических связей)
- D. Программное извлечение из текста на РЯ (на основе мер ассоциаций) наиболее частотных неразрывных двухсловных коллокаций и их анализ
- E. Автоматическое извлечение терминов из текста на основе комбинации признаков, оценка средней точности

Срок выполнения задания – до **10 апреля** включительно

# ЛОКАЛЬНЫЕ ВЫСОКОВЕРОЯТНЫЕ СИНТАКСИЧЕСКИЕ СВЯЗИ В РЯ



Из наиболее распространенных:

- *V* и *N* (вин. падеж): *перевозит* → *грузы*
- *N* и *N* (род. падеж): *перевозка* → *грузов*,
- *N* и *A* (согласованные): *интересная* ← *книга*
- *P* и *A* (согласованные): *прочитанная* ← *книга*
- *V* и *V* (инфинитив): *умеет* → *плавать*
- *N* и *V* (инфинитив): *умение* → *плавать*
- *A* и *V* (инфинитив): *готовый* → *помочь*
- *Adv* и *Adv*: *очень* ← *хорошо*
- *Adv* и *A*: *весьма* ← *интересный*
- *Adv* и *V*: *быстро* ← *бежит*
- *Num* и *N*: *пять* ← *машин*
- ...

# ПРАВИЛО УСТАНОВЛЕНИЯ ЛОКАЛЬНОЙ СВЯЗИ ПРЕДЛОГА



Правило установления локальной зависимости

$Prep \rightarrow N$  (предлог-существительное): *в город*

- Если падеж  $N$  соответствует падежам, обслуживаемым предлогом  $Prep$ , то установить связь, сделав  $Prep$  главным словом
- Если  $N$  является неизменяемым, то установить связь, сделав  $Prep$  главным словом
- В иных случаях связь не устанавливать

Правило установления зависимости :  $N \rightarrow Prep$

(существительное-предлог): *освобождение от*

- Если  $N$  является отглагольным, то сделать его главным и установить связь
- В противном случае связь не устанавливать