

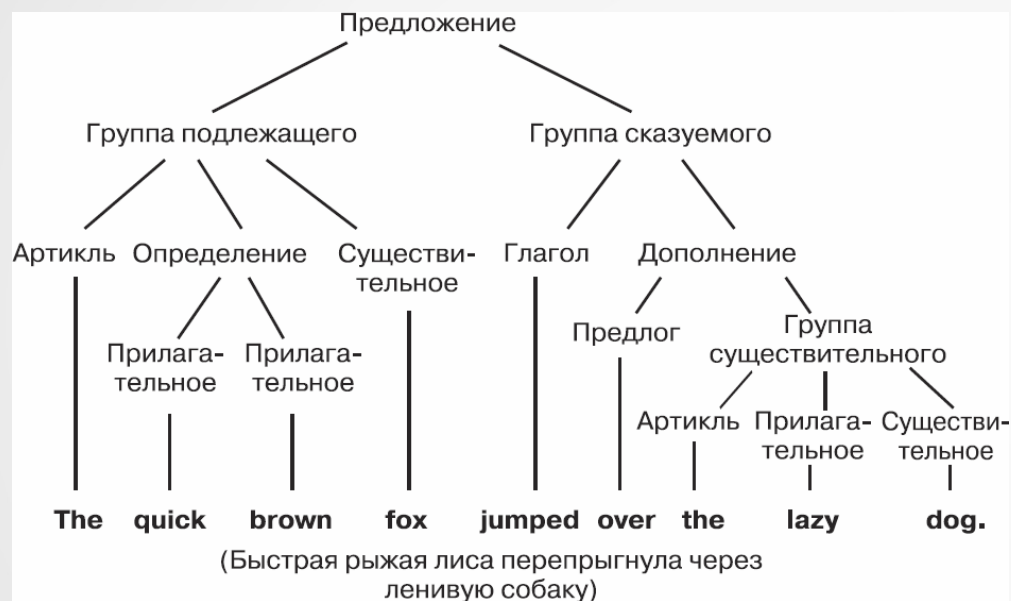


# **Формальные грамматики**

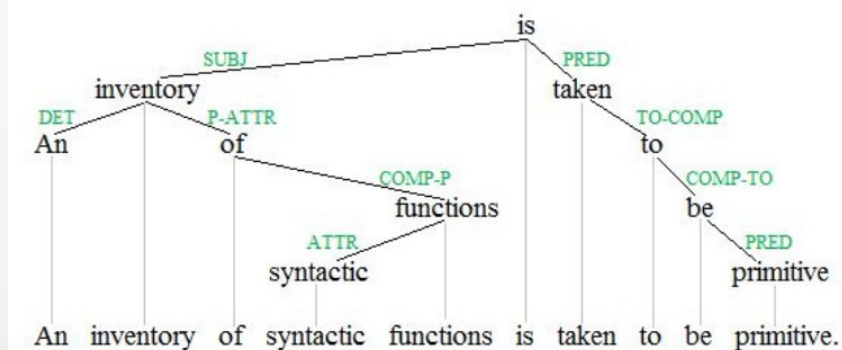
Евгений Борисов

# Синтаксический анализ

## грамматика составляющих



## грамматика зависимостей



# Синтаксический анализ

## грамматика составляющих (constituency grammar)

— разметка (вложенных) групп

именная группа (группа существительного, ИГ; англ. noun phrase, NP) — возглавляется существительным;

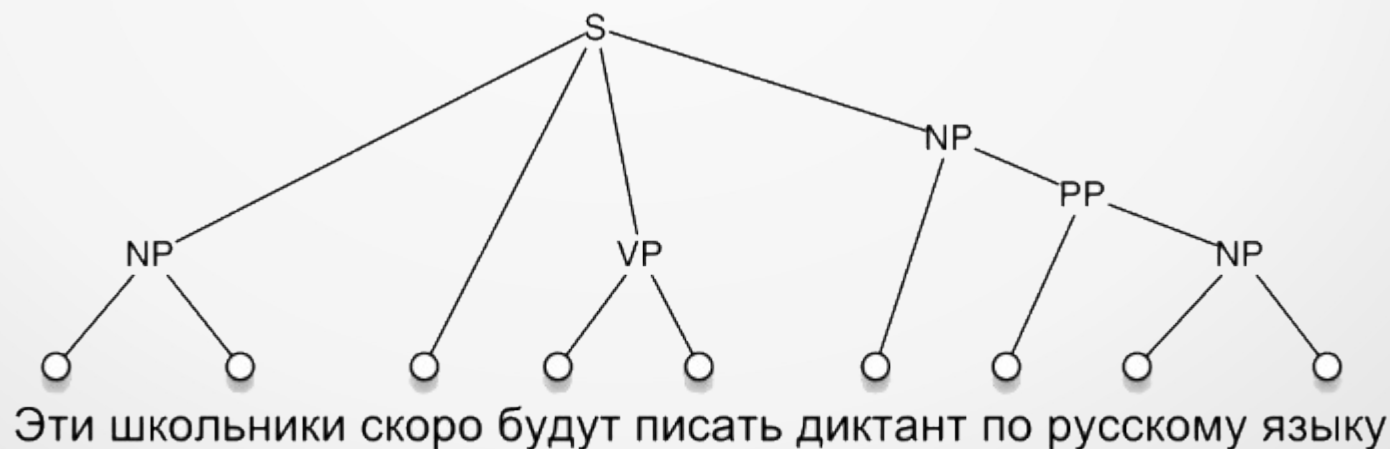
группа прилагательного (ГПрил; adjectival phrase, AP) — возглавляется прилагательным;

наречная группа (НарГ; adverbial phrase, AdvP) — возглавляется наречием;

предложная группа (ПрГ; prepositional phrase, PP) — возглавляется предлогом;

глагольная группа (ГГ; verb phrase, VP) — возглавляется глаголом;

предложение (П; sentence, S).



[s[NP Эти школьники] скоро[VP будут писать][NP диктант[PP по [NP русскому языку]]]]

# Синтаксический анализ

**Формальная грамматика** - способ описания языка

$$G=(N,\Sigma,R,s); V=N\cup\Sigma$$

$N$  — множество (алфавит) нетерминальных символов (синтаксические переменные или понятия)

$\Sigma$  - множество (алфавит) терминальных символов (не пересекается с  $N$ )

$V$  - словарь грамматики  $G$

$s$  - начальный нетерминал (принадлежит алфавиту нетерминалов  $N$ )

$R$  - конечное множество правил вывода (продукции),  
вида  $A \rightarrow b$ , где  $A, b$  — последовательности символов из алфавита  $V$  грамматики  $G$

## **Нетерминальные символы**

- объекты, обозначающие какую-либо сущность языка (предложение, формула и т.д.).

## **Терминальные символы**

- объекты непосредственно присутствующие в языке.

# Синтаксический анализ

**Форма Бэкуса-Наура (БНФ)** - способ представления КС-грамматик

**Lex/Yacc калькулятор:**

```
%token INTEGER
```

```
expr: INTEGER  
    | "-" expr  
    | "(" expr ")"  
    | expr "-" expr  
    | expr "+" expr  
    | expr "*" expr  
    | expr "/" expr
```

**expr** — нетерминальный символ (объект, обозначающий сущность языка)

**- + ( ) \* / число** — терминальные символы (объекты непосредственно присутствующие в языке)

(пример описывает только INT)

# Синтаксический анализ

## классификация формальных грамматик по Хомскому

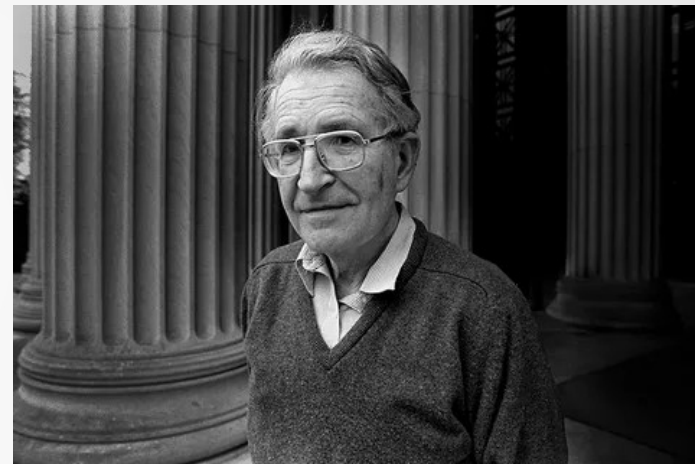
неограниченные

$R : V \rightarrow \beta$

$V$  — любая непустая последовательность из  $V$  содержащая нетерминалы

$\beta$  — любая (в т.ч. пустая) последовательность из  $V$

$$G = (N, \Sigma, R, s); V = N \cup \Sigma$$



*Avram Noam Chomsky*

# Синтаксический анализ

## классификация формальных грамматик по Хомскому

### неограниченные

$$R : B \rightarrow \beta$$

$B$  — любая непустая последовательность из  $V$  содержащая нетерминалы

$\beta$  — любая (в т.ч. пустая) последовательность из  $V$

### контекстно-зависимые

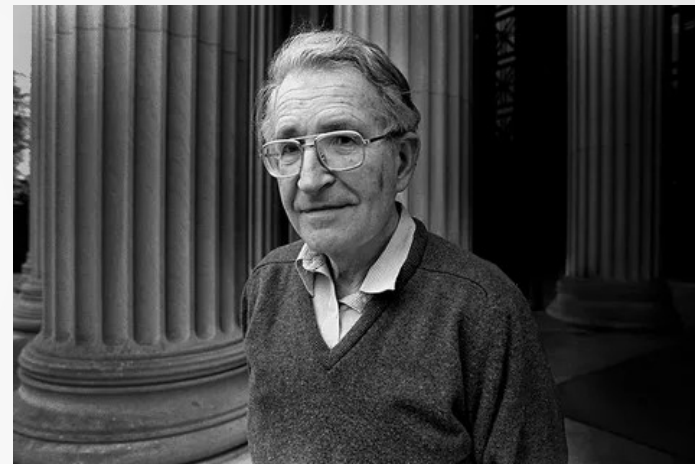
$$R : \alpha A \beta \rightarrow \alpha \gamma \beta$$

$A$  — нетерминал из  $N$

$\alpha, \beta$  — любые (в т.ч. пустые) последовательности из  $V$

$\gamma$  — любая непустая последовательность из  $V$

$$G = (N, \Sigma, R, s); V = N \cup \Sigma$$



*Avram Noam Chomsky*

# Синтаксический анализ

## классификация формальных грамматик по Хомскому

### неограниченные

$R : V \rightarrow \beta$

$V$  — любая непустая последовательность из  $V$  содержащая нетерминалы  
 $\beta$  — любая (в т.ч. пустая) последовательность из  $V$

### контекстно-зависимые

$R : \alpha A \beta \rightarrow \alpha \gamma \beta$

$A$  — нетерминал из  $N$   
 $\alpha, \beta$  — любые (в т.ч. пустые) последовательности из  $V$   
 $\gamma$  — любая непустая последовательность из  $V$

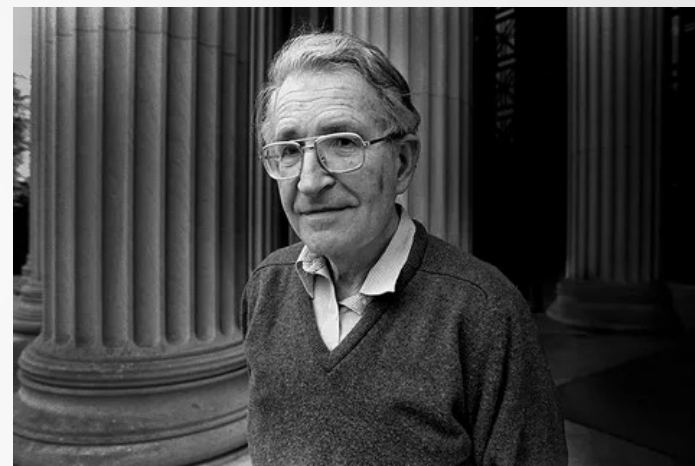
### контекстно-свободные

$R : A \rightarrow \beta$

$A$  — нетерминал из  $N$   
 $\beta$  — любая (в т.ч. пустая) последовательность из  $V$

*применяются для описания компьютерных языков*

$$G = (N, \Sigma, R, s); V = N \cup \Sigma$$



*Avram Noam Chomsky*



# Синтаксический анализ

## классификация формальных грамматик по Хомскому

### неограниченные

$$R : B \rightarrow \beta$$

$B$  — любая непустая последовательность из  $V$  содержащая нетерминалы  
 $\beta$  — любая (в т.ч. пустая) последовательность из  $V$

### контекстно-зависимые

$$R : \alpha A \beta \rightarrow \alpha \gamma \beta$$

$A$  — нетерминал из  $N$   
 $\alpha, \beta$  — любые (в т.ч. пустые) последовательности из  $V$   
 $\gamma$  — любая непустая последовательность из  $V$

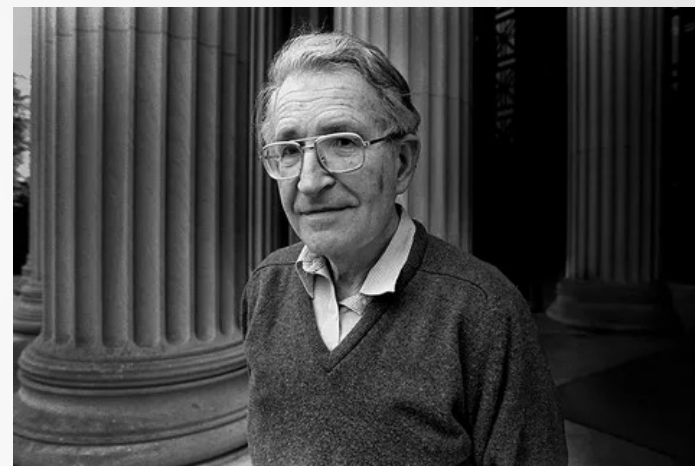
### контекстно-свободные

$$R : A \rightarrow \beta$$

$A$  — нетерминал из  $N$   
 $\beta$  — любая (в т.ч. пустая) последовательность из  $V$

*применяются для описания компьютерных языков*

$$G = (N, \Sigma, R, s); V = N \cup \Sigma$$



Avram Noam Chomsky

### регулярные

$$R : A \rightarrow A\beta, A \rightarrow \beta \text{ (леворекурсивные)}$$

$$R : A \rightarrow \beta A, A \rightarrow \beta \text{ (праворекурсивные)}$$

$A$  — нетерминал из  $N$   
 $\beta$  — последовательность (в т.ч. пустая) терминалов из  $\Sigma$

*применяются для описания простых конструкций*

# Синтаксический анализ

## Нормальная форма грамматики по Хомскому (CNF)

продукции имеют вид:  $A \rightarrow BC$ ,  $A \rightarrow \alpha$ ,  $s \rightarrow \epsilon$ ,

где

$A, B, C$  – нетерминалы ( $B$  и  $C$  не могут являться начальными символами),

$\alpha$  – терминальный символ,

$s$  – начальный символ,

$\epsilon$  – пустая строка.

## Эквивалентность грамматик

- сильная (совпадает язык и дерево разбора)
- слабая (совпадает язык, деревья разбора могут отличаться)

## Теорема:

любая КС-грамматика может быть преобразована в эквивалентную CNF-грамматику.

# Синтаксический анализ

## Контекстно-свободные грамматики и регулярные языки

- Контекстно-свободные грамматики являются обобщением регулярных грамматик
- Центральная вставка  $A \rightarrow \alpha A \beta$
- Пример:
  - The luggage arrived.
  - The luggage that the passengers checked arrived.
  - The luggage that the passengers that the storm delayed checked arrived.

# Синтаксический анализ

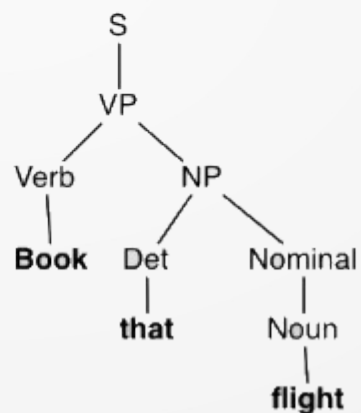
## **Методы синтаксического разбора**

- рекурсивный спуск (top-down parsing)
- восходящий анализ (bottom-up parsing)
- алгоритм Кока-Янгера-Касами (CKY parsing)
- алгоритм Эрли (Earley parser)
- ...

# Синтаксический анализ

S → NP VP  
S → Aux NP VP  
S → VP  
NP → Pronoun  
NP → Proper-Noun  
NP → Det Nominal  
Nominal → Noun  
Nominal → Nominal Noun  
Nominal → Nominal PP  
VP → Verb  
VP → Verb NP  
VP → Verb NP PP  
VP → Verb PP  
VP → VP PP  
PP → Preposition NP

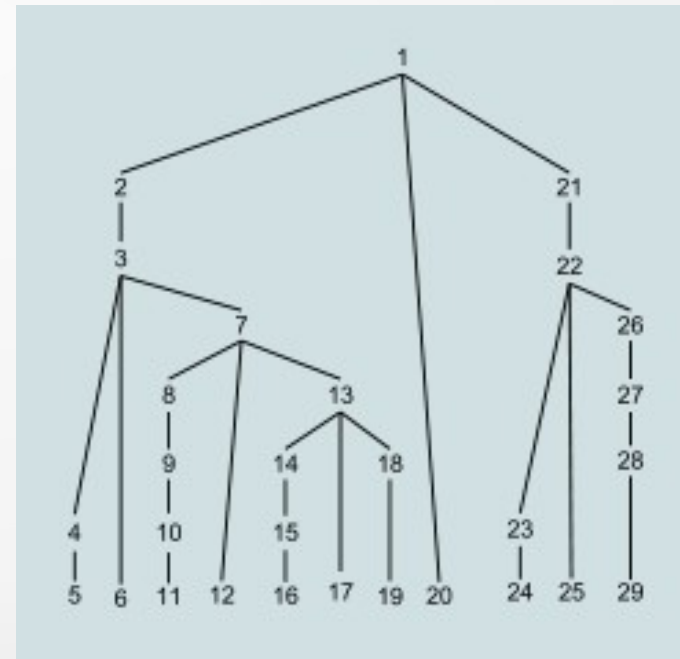
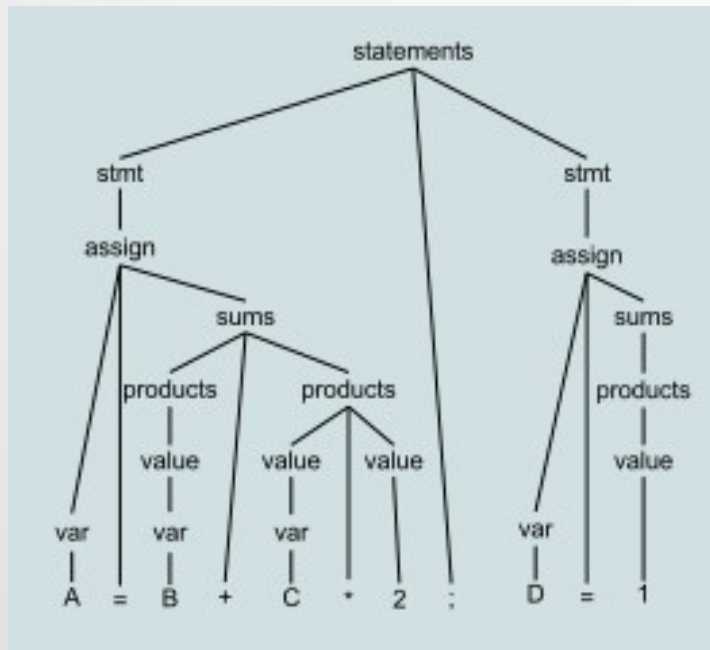
Det → that | this | a  
Noun → book | flight | meal | money  
Verb → book | include | prefer  
Pronoun → I | she | me  
Proper-Noun → Houston | TWA  
Aux → does  
Preposition → from | to | on | near | through



# Синтаксический анализ

## Анализ рекурсивным спуском (top-down parsing)

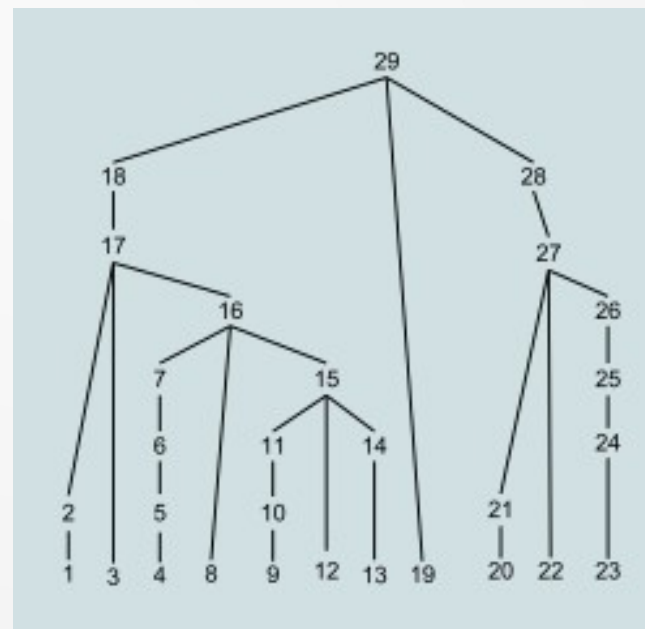
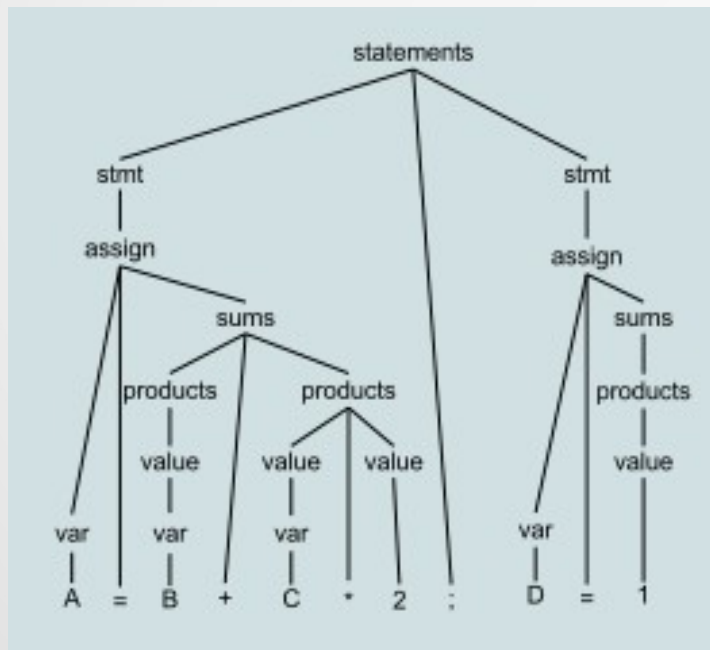
правила формальной грамматики раскрываются, начиная со стартового символа, до получения требуемой последовательности токенов.



# Синтаксический анализ

## Восходящий анализ (bottom-up parsing)

сначала распознает мелкие детали самого низкого уровня текста, а затем его структуры среднего уровня, и оставляет общую структуру самого высокого уровня на потом.



# Синтаксический анализ

## Частичный разбор (группировка, применение группы правил)

Применяется для извлечения именованных сущностей

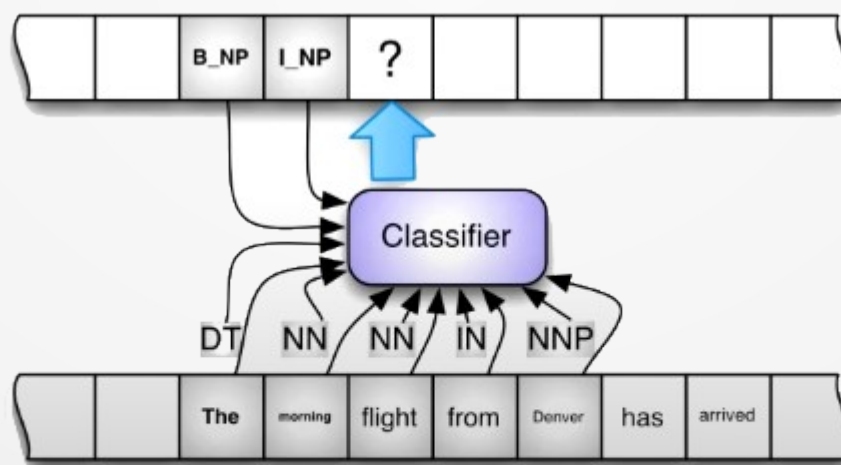
- Partial parsing, Shallow parsing
- Chunking, фрагментирование
  - <sub>[NP</sub> The morning flight]<sub>[PP</sub> from]<sub>[NP</sub> Denver]<sub>[VP</sub> has arrived]
  - <sub>[NP</sub> The morning flight] from <sub>[NP</sub> Denver] has arrived



# Синтаксический анализ

## Группировка на основе машинного обучения

- Классы BIO (begin, inside, outside)
- Тренировочное множество - Treebank



Признаки: *The*, *DT*, *B\_NP*, *morning*, *NN*, *I\_NP*, *flight*, *NN*, *from*, *IN*, *Denver*, *NNP*

# Синтаксический анализ

## Статистические КС-грамматики

$$G=(N,\Sigma,R,s); V=N\cup\Sigma$$

$N$  — множество (алфавит) нетерминальных символов (синтаксические переменные или понятия)

$\Sigma$  - множество (алфавит) терминальных символов (не пересекается с  $N$ )

$V$  - словарь грамматики  $G$

$s$  - начальный нетерминал (принадлежит алфавиту нетерминалов  $N$ )

$R$  - конечное множество правил вывода (продукции),

вида  $A \rightarrow \beta[p]$

где

$A$  — нетерминал из  $N$

$\beta$  — последовательности символов из алфавита  $V$  грамматики  $G$

$p$  — **вероятность правила**  $P(\beta|A)$  (сумма вероятностей всех правил вида  $A \rightarrow *$  равна 1)

### Нетерминальные символы

- объекты, обозначающие какую-либо сущность языка (предложение, формула и т.д.).

### Терминальные символы

- объекты непосредственно присутствующие в языке.

# Синтаксический анализ

## Статистические КС-грамматики

Грамматика	Вероятность	Лексикон
$S \rightarrow NP VP$	0.8	$Det \rightarrow the \mid a \mid that \mid this$
$S \rightarrow Aux NP VP$	0.1	0.6 0.2 0.1 0.1
$S \rightarrow VP$	0.1	$Noun \rightarrow book \mid flight \mid meal \mid money$
$NP \rightarrow Pronoun$	0.2	0.1 0.5 0.2 0.2
$NP \rightarrow Proper-Noun$	0.2	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Det Nominal$	0.6	0.5 0.2 0.3
$Nominal \rightarrow Noun$	0.3	$Pronoun \rightarrow I \mid he \mid she \mid me$
$Nominal \rightarrow Nominal Noun$	0.2	0.5 0.1 0.1 0.3
$Nominal \rightarrow Nominal PP$	0.5	$Proper-Noun \rightarrow Houston \mid NWA$
$VP \rightarrow Verb$	0.2	0.8 0.2
$VP \rightarrow Verb NP$	0.5	$Aux \rightarrow does$
$VP \rightarrow VP PP$	0.3	1.0
$PP \rightarrow Prep NP$	1.0	$Prep \rightarrow from \mid to \mid on \mid near \mid through$
		0.25 0.25 0.1 0.2 0.2

# Синтаксический анализ

## Статистические КС-грамматики

### Разрешение многозначности

- Вероятность разбора

$$P(T, S) = \prod_{i=1}^n P(RHS_i | LHS_i)$$

- Вероятность  $P(T, S) = P(T)P(S|T) = P(T)$

- Выбор наиболее вероятного дерева разбора  $\hat{T}(S) = \arg \max_T P(T|S)$

$$\hat{T}(S) = \arg \max_T \frac{P(T, S)}{P(S)}$$

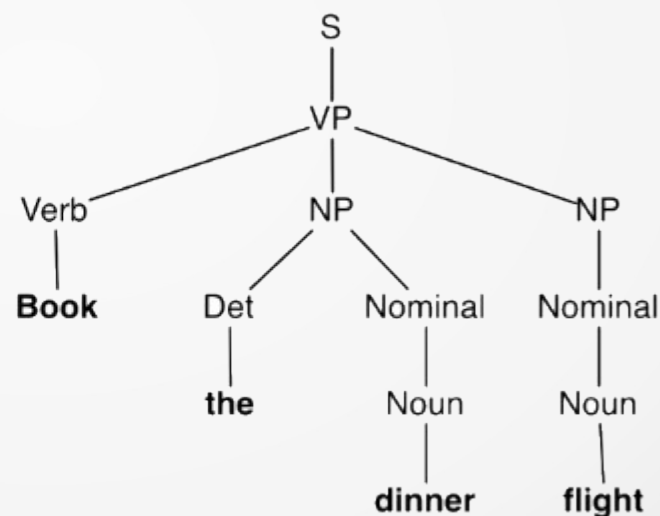
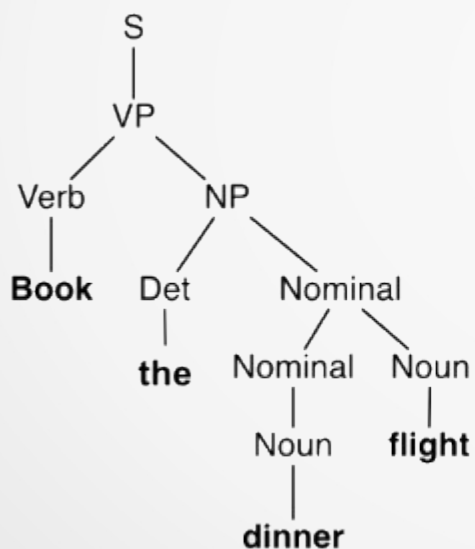
$$\hat{T}(S) = \arg \max_T P(T, S)$$

$$\hat{T}(S) = \arg \max_T P(T)$$

# Синтаксический анализ

## Статистические КС-грамматики

### Разрешение многозначности



$$P(\text{T-left}) = .05 \cdot .20 \cdot .20 \cdot .20 \cdot .75 \cdot .30 \cdot .60 \cdot .10 \cdot .40 = 2.2 \cdot 10^{-6}$$

$$P(\text{T-right}) = .05 \cdot .10 \cdot .20 \cdot .15 \cdot .75 \cdot .75 \cdot .30 \cdot .60 \cdot .10 \cdot .40 = 6.1 \cdot 10^{-7}$$

# Синтаксический анализ

## Статистические КС-грамматики

### Обучение КС

- Вычисление вероятности на основе банка деревьев

$$P(\alpha \rightarrow \beta | \alpha) = \frac{\text{Count}(\alpha \rightarrow \beta)}{\sum_{\gamma} \text{Count}(\alpha \rightarrow \gamma)} = \frac{\text{Count}(\alpha \rightarrow \beta)}{\text{Count}(\alpha)}$$

- Вывод без тренировочного множества (ЕМ)
  - На основе множества предложений построить множество наиболее вероятных синтаксических разборов
  - Обновить значения вероятностей на основе полученных данных
  - (Manning and Schutze 1999)

# Синтаксический анализ

## Оценка качества алгоритма

- Метрика PARSEVAL: пусть  $P$  - дерево разбора, созданное алгоритмом,  $T$  - дерево разбора, созданное экспертами
  - Точность =  $(\# \text{ правильных компонент в } P) / (\# \text{ компонент в } T)$
  - Полнота =  $(\# \text{ правильных компонент в } P) / (\# \text{ компонент в } P)$
  - F-мера =  $2PR / (P + R)$
- Современные алгоритмы показывают точность и полноту более 90%

# Синтаксический анализ

## Литература

git clone [https://github.com/mechanoid5/ml\\_nlp.git](https://github.com/mechanoid5/ml_nlp.git)

Турдаков Д.Ю.

Основы обработки текстов. лекция 7.

Формальные грамматики и синтаксический анализ. ИСП РАН, 2017

<https://www.youtube.com/watch?v=TkMtUm-D6aE>

Steven Bird, Ewan Klein, and Edward Loper  
Analyzing Text with the Natural Language Toolkit

<https://www.nltk.org/book/>

D.Jurafsky, J.H.Martin Speech and Language Processing. third edition, 2020

А. Ахо, Дж. Ульман. Теория синтаксического анализа, перевода и компиляции. М.: Мир, 1978.

Д.Кук, Г.Бейз Компьютерная математика - Москва: Наука, 1990

В.С.Проценко, П.Й.Чаленко Элементы компиляции - Киев: УМК ВО, 1988

Е.С.Борисов Методы и средства построения грамматических анализаторов.

<http://mechanoid.su/programming-grammar-analysis.html>