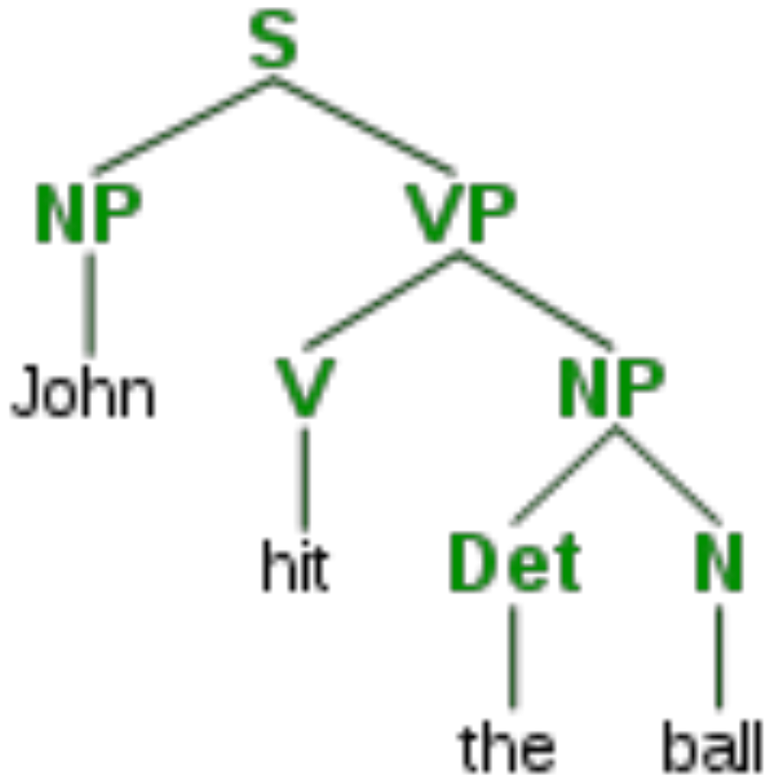
# Основы обработки текстов

Лекция 6

Формальные грамматики и синтаксический анализ

# Пример синтаксического разбора



# Где может быть полезно знание синтаксиса?

- Машинный перевод
- Генерация текста
  - –диалоговые системы
- Извлечение информации
- Понимание на что/кого направлено эмоциональное высказывание
- ...

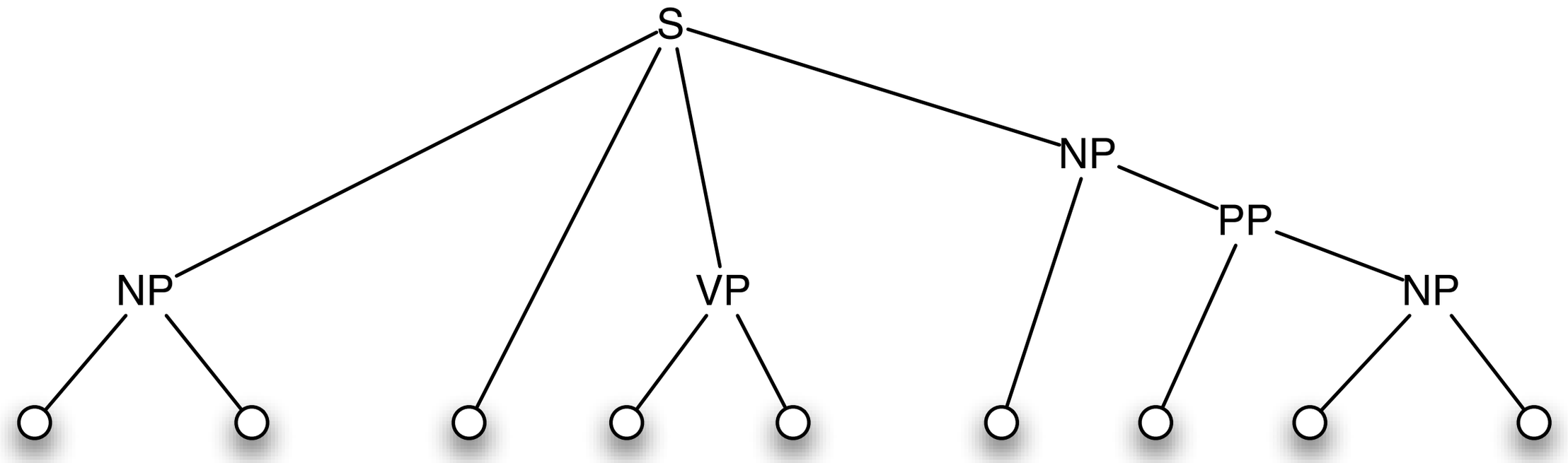
# План

- Грамматика естественного языка
- Формальные грамматики
  - –Контекстно-свободные грамматики
  - –Грамматики зависимостей
  - –Категориальные грамматики
- Синтаксический разбор
- Группировка (Фрагментирование)

# Грамматика составляющих

- именная группа (группа существительного, noun phrase, NP)
- группа прилагательного (adjectival phrase, ADJP)
- наречная группа (adverbial phrase, ADVP)
- предложная группа (prepositional phrase, PP)
- глагольная группа (verb phrase, VP);

# Пример




Эти школьники скоро будут писать диктант по русскому языку

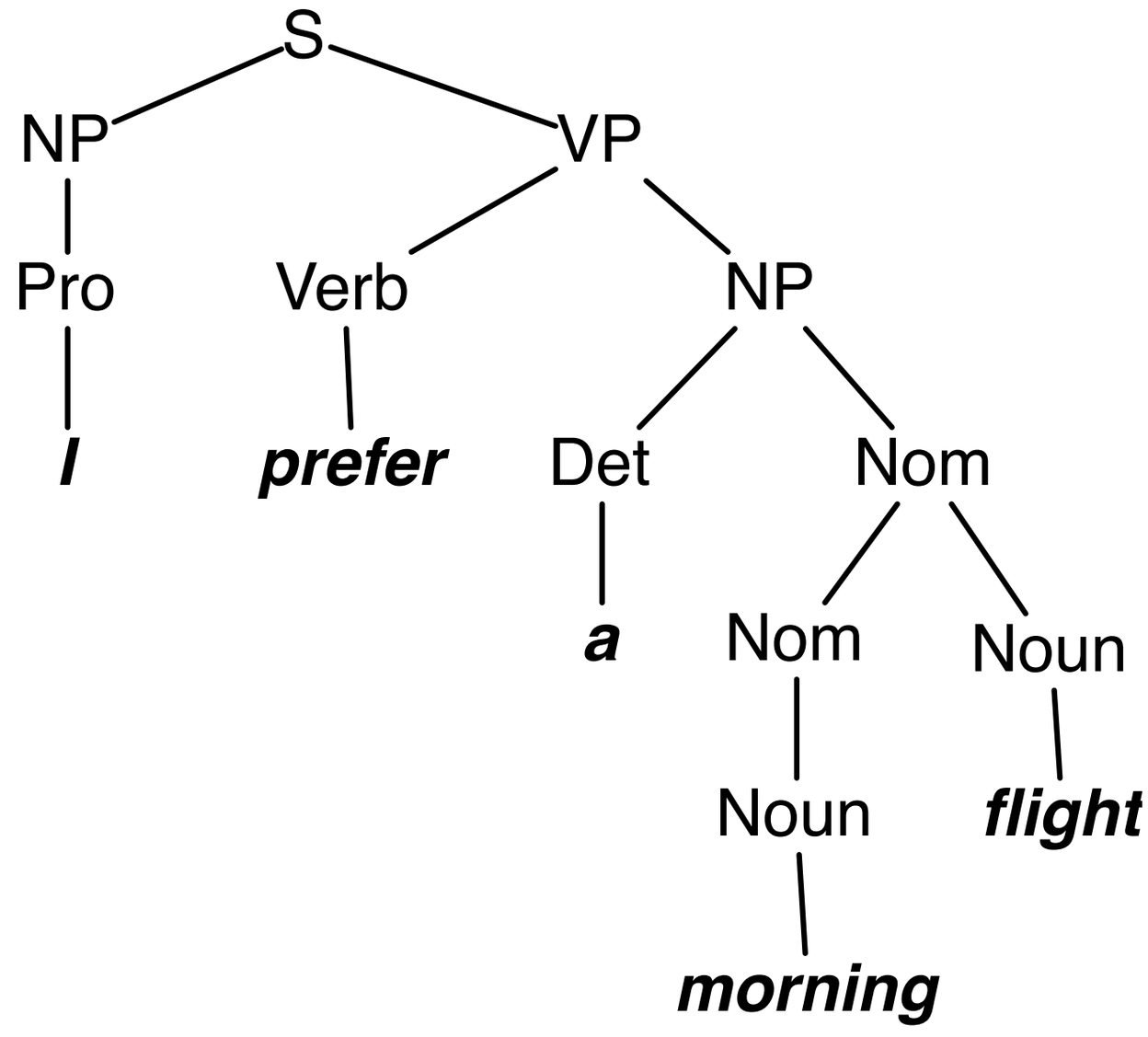
[s[np Эти школьники] скоро[vp будут писать][np диктант[pp по [np русскому языку]]]]

# Контекстно свободные грамматики

| | | |
|---|---|---|
| Noun | → | flights \| breeze \| trip \| morning |
| Verb | → | is \| prefer \| like \| need \| want\| fly |
| Adjective | → | cheapest \| non-stop \| first \| latest \| other \| direct |
| Pronoun | → | me \| I \| you \| it |
| Proper-Noun | → | Alaska \| Los Angeles \| Chicago |
| Determiner | → | the \| a \| an \| this \| these \| that |
| Preposition | → | from \| to \| on \| near |
| Conjunction | → | and \| or \| but |

| | | | |
|---|---|---|---|
| S | → | NP VP | I + want a morning flight |
| NP | → | Pronoun | I |
| | \| | Proper-Noun | Los Angeles |
| | \| | Det Nominal | a + flight |
| Nominal | → | Nominal Noun | morning + flight |
| | \| | Noun | flights |
| VP | → | Verb | do |
| | \| | Verb NP | want + a flight |
| | \| | Verb NP PP | leave + Boston + in the morning |
| | \| | Verb PP | leaving + on Thursday |
| PP | → | Preposition NP | from + Los Angeles |

# Пример


S
NP
VP
Pro
Verb
NP
I
prefer
Det
Nom
a
Nom
Noun
Noun
flight
morning

# Формальное определение

| | |
|---|---|
| $N$ | множество **нетерминальных символов** |
| $\Sigma$ | множество **терминальных символов** (непересекающееся с *N*) |
| $R$ | множество правил, каждое вида $A \to \beta$ где $A$ - нетерминал, $\beta$ - строка символов из множества $(\Sigma \cup N)*$ |
| $S$ | символ начала |

# Согласование

- Пример
  - –по русскому языку
  - –русский язык
- **Проблема:** Увеличение количества правил
- **Решение:** Введение параметров для нетерминальных символов
  - –см. Jurafsky, Martin глава 15

# Откуда взять грамматику?

- Написать вручную
- Вывод грамматики по банку деревьев
  - Penn Treebank Project

```
( (S (NP-SBJ (NP Pierre Vinken)
             ,
             (ADJP (NP 61 years)
                                 old)
             ,)
     (VP will
        (VP join
           (NP the board)
           (PP-CLR as
                                    (NP a nonexecutive director))
                       (NP-TMP Nov. 29)))
     .))
( (S (NP-SBJ Mr. Vinken)
     (VP is
        (NP-PRD (NP chairman)
                                 (PP of
                                    (NP (NP Elsevier N.V.)
                                                   ,
                                                   (NP the Dutch publishing group)))))
     .))
```

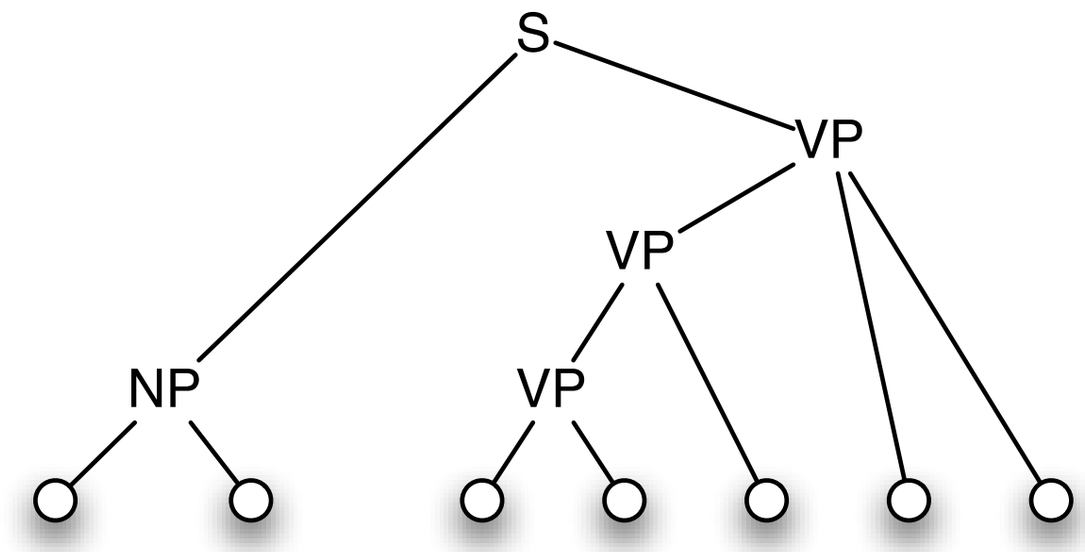
# Эквивалентность грамматик

- Эквивалентность
  - –сильная (язык + деревья разбора)
  - –слабая (только язык)
- Нормальная форма грамматики (Хомского)
  - – $A \rightarrow BC$
  - – $A \rightarrow a$
- Всегда существует преобразование в нормальную форму (слабая эквивалентность)

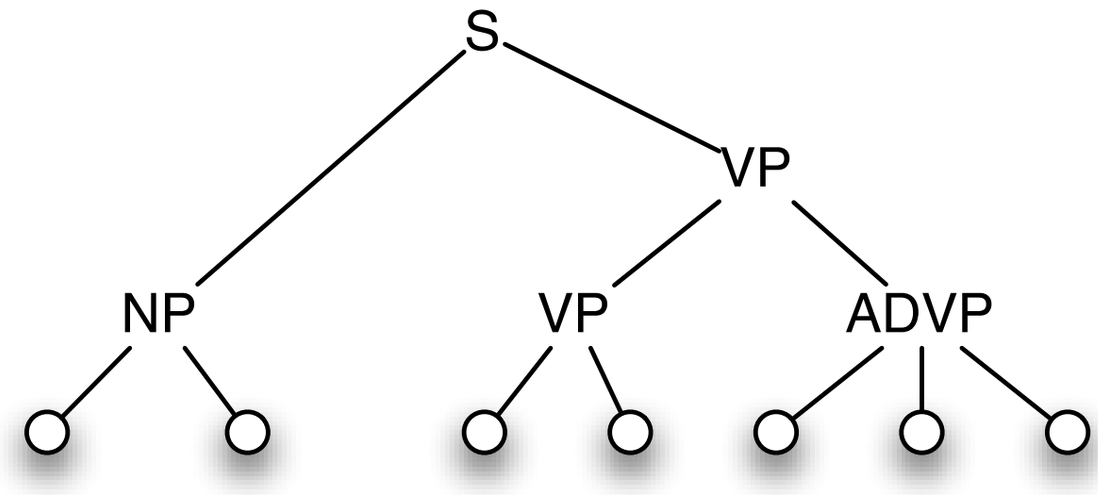
# Контекстно-свободные грамматики и регулярные языки

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

# Синтаксическая многозначность




Народ Беларуси будет жить плохо, но недолго (А.Г. Лукашенко)

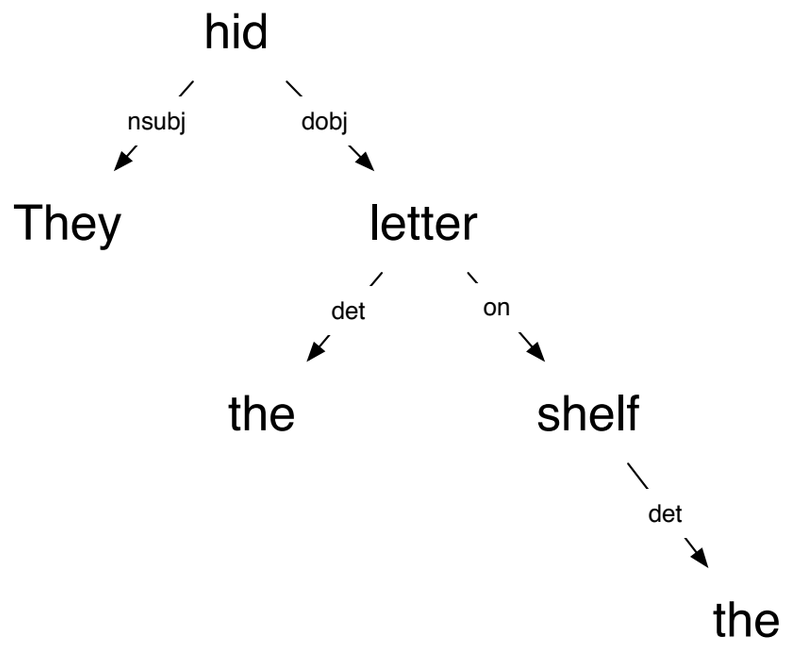



Народ Беларуси будет жить плохо, но недолго (А.Г. Лукашенко)

# Другие типы грамматик

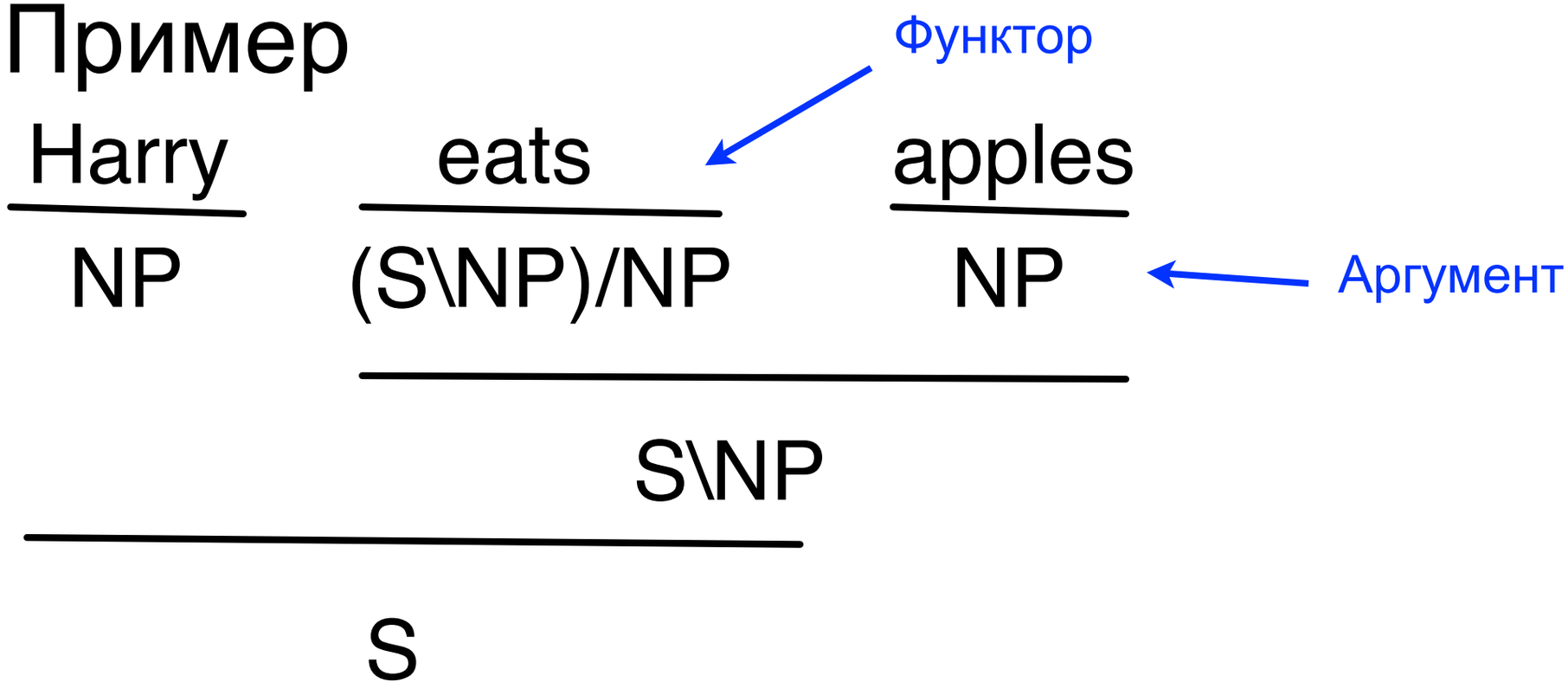
# Грамматика зависимостей

- Способность предсказывать аргументы при синтаксическом разборе
- Хорошо отражают специфику языков с произвольным порядком слов
- Может быть автоматически получена из дерева разбора на составляющие



# Категориальная грамматика

- Категории фраз:
  - –Состоят из функторов и аргуменов
  - –X/Y - функция из Y в X. Аргумент присоединяется к Y справа, чтобы получилось X
  - –X\Y - ... слева ...
- Пример



# Синтаксический разбор

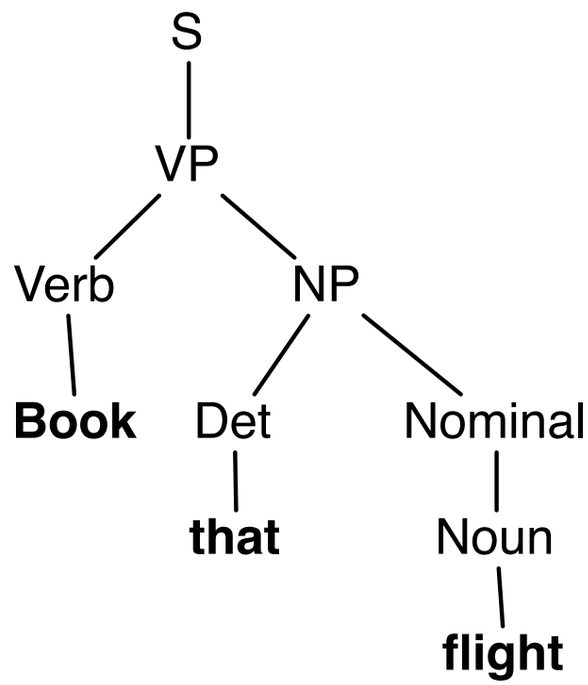
# Синтаксический разбор

- Рассматриваемые алгоритмы
  - –Метод рекурсивного спуска (top-down parsing)
  - –Восходящий анализ (bottom-up parsing)
  - –Алгоритм Кока-Янгера-Касами (CKY Parsing)
- Не рассматриваемые, но часто используемые алгоритмы
  - –Алгоритм Эрли (Earley parser)
  - –Chart parser
  - –http://en.wikipedia.org/wiki/Category:Parsing_algorithms

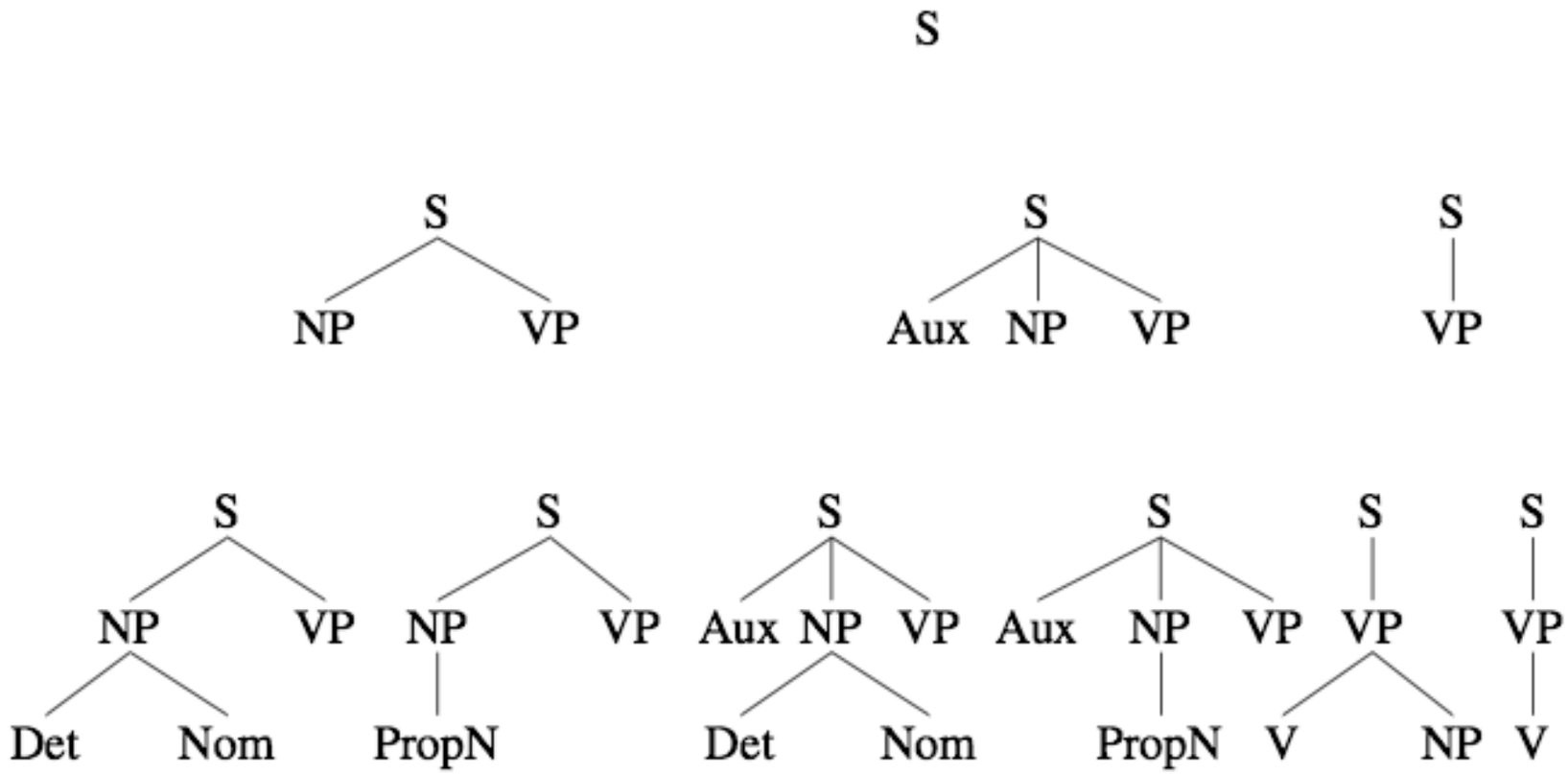
# Пример

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

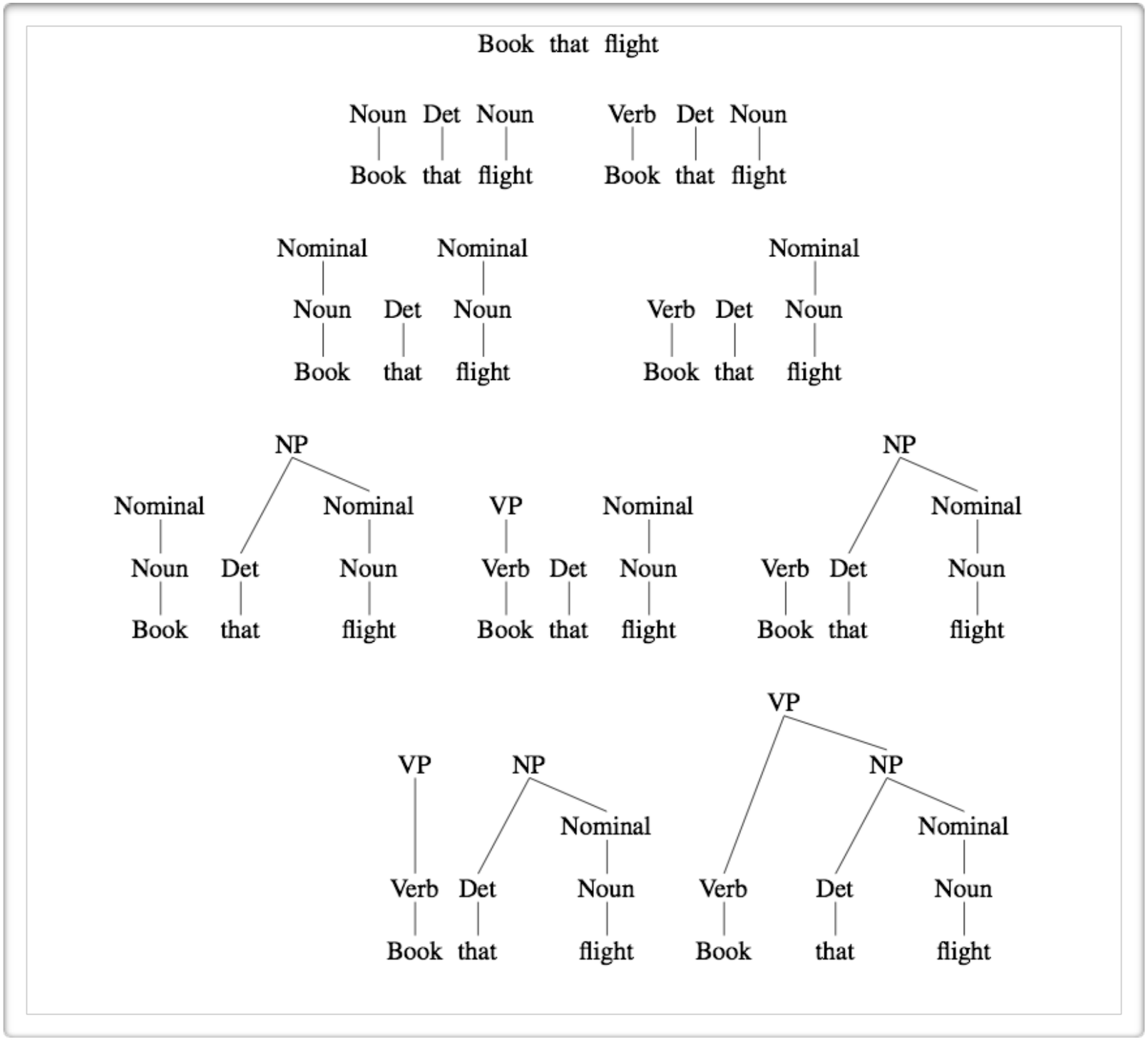
Det → that | this | a
Noun → book | flight | meal | money
Verb → book | include | prefer
Pronoun → I | she | me
Proper-Noun → Houston | TWA
Aux → does
Preposition → from | to | on | near | through



# Метод рекурсивного спуска



# Восходящий анализ



# Алгоритм CKY

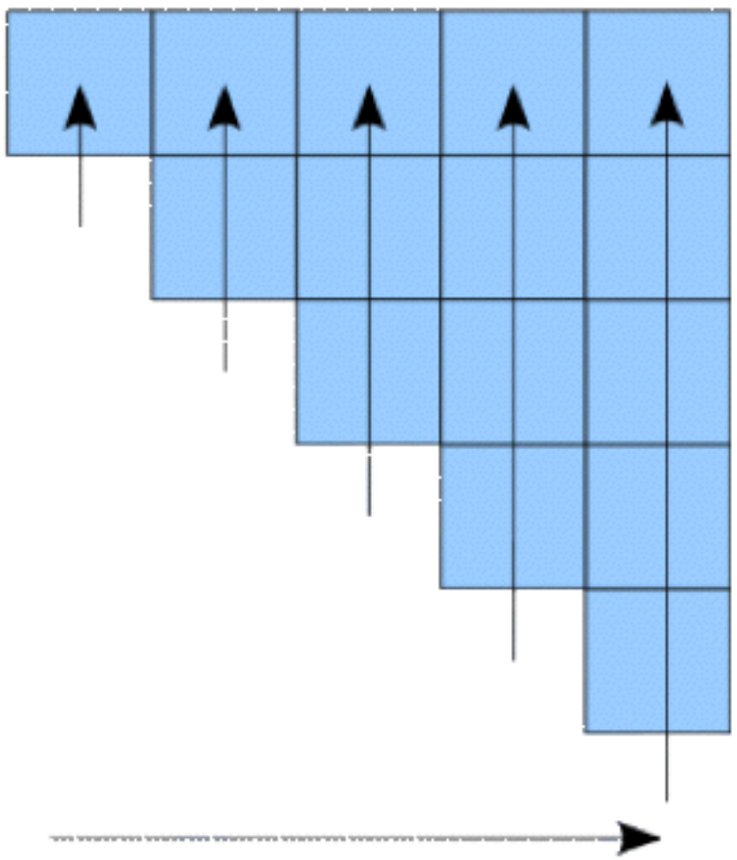
- Шаг 0. Преобразовать грамматику к нормальной форме
- Алгоритм (динамическое программирование)

**function** CKY-PARSE(*words, grammar*) **returns** *table*

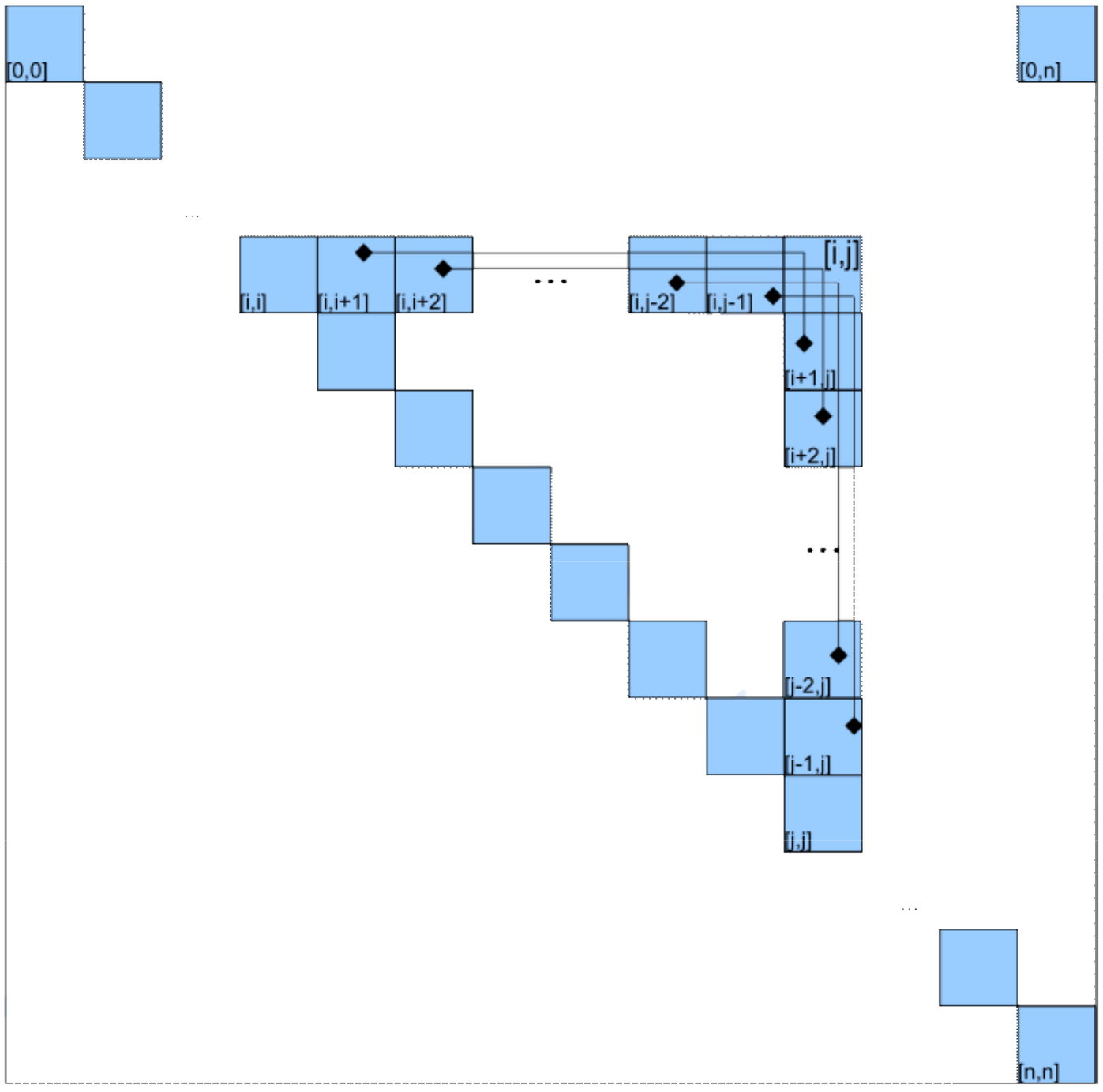
**for** $j \leftarrow$ **from** 1 **to** LENGTH(*words*) **do**
  $table[j-1, j] \leftarrow \{A \mid A \rightarrow words[j] \in grammar\}$
  **for** $i \leftarrow$ **from** $j-2$ **downto** 0 **do**
    **for** $k \leftarrow i+1$ **to** $j-1$ **do**
      $table[i,j] \leftarrow table[i,j] \cup$
        $\{A \mid A \rightarrow BC \in grammar,$
          $B \in table[i,k],$
          $C \in table[k,j]\}$

# Распознавание

| Book | the | flight | through | Houston |
|---|---|---|---|---|
| SP, VP, Nominal, Verb,Noun [0,1] | [0,2] | S, VP, X2 [0,3] | [0,4] | S1, VP1, S2, VP2, S3 [0,5] |
| | Det [1,2] | NP [1,3] | [1,4] | NP [1,5] |
| | | Nominal, Noun [2,3] | [2,4] | Nominal [2,5] |
| | | | Prep [3,4] | PP [3,5] |
| | | | | NP, Proper-Noun [0,1] |

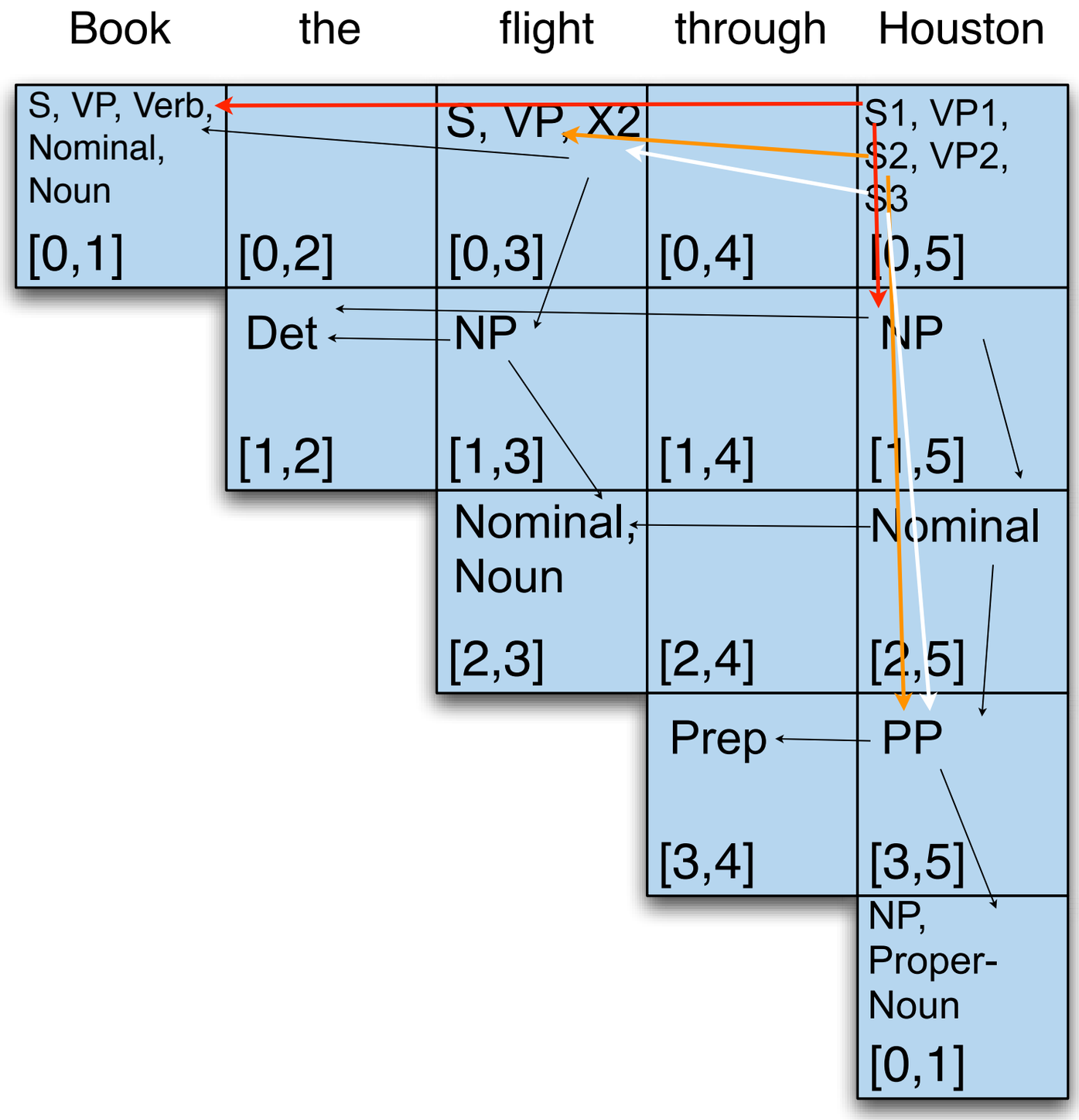


# Запоминание путей



# Синтаксический разбор

S → NP VP
S → X1 VP
X1 → Aux NP
S → VP
S → X2 PP
NP → Pronoun
NP → Proper-Noun
NP → Det Nominal
Nominal → Noun
Nominal → Nominal Noun
Nominal → Nominal PP
VP → Verb
VP → Verb NP
VP → X2 PP
X2 → Verb NP
VP → Verb PP
VP → VP PP
PP → Preposition NP



# Группировка

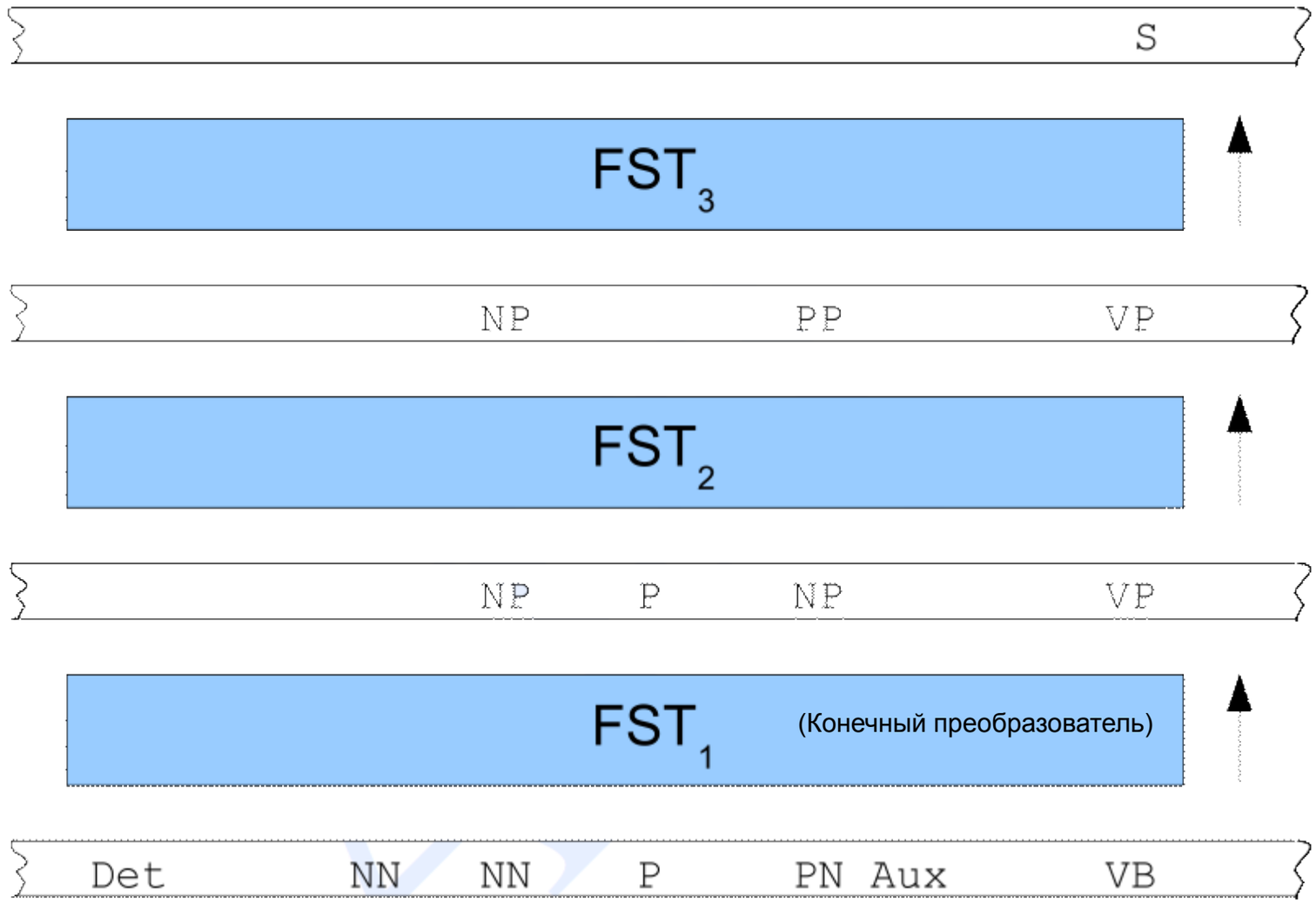
- Partial parsing, Shallow parsing
- Chunking, фрагментирование
  - –[NP The morning flight][PP from][NP Denver][VP has arrived]
  - –[NP The morning flight] from [NP Denver] has arrived

# Группировка на основе правил

S → PP* NP PP* VP PP*

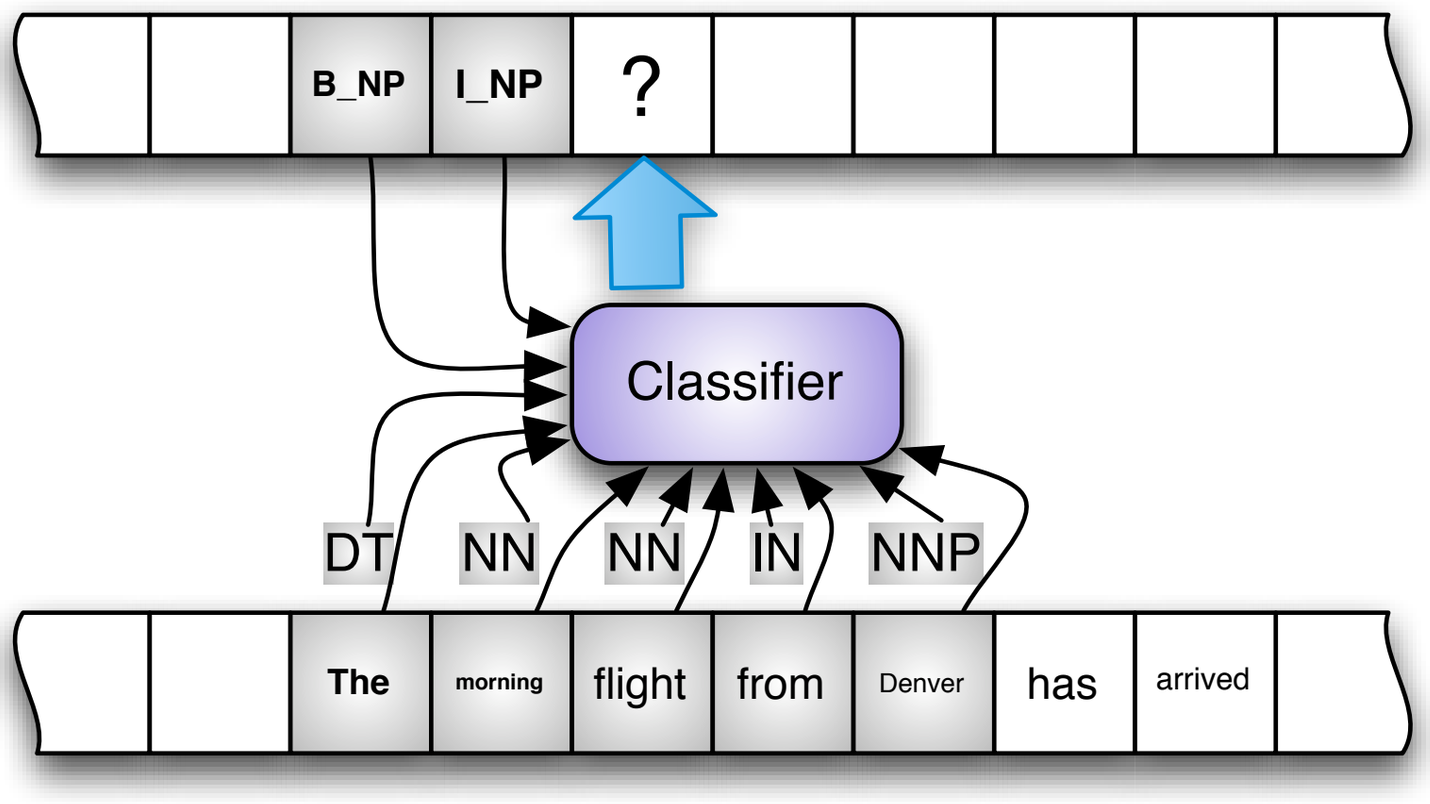
PP → IN NP

NP → (Det) Noun* Noun
NP → Proper-Noun
VP → VB
VP → Aux VB



# 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*

# Заключение

- Изучены
  - –некоторые особенности грамматик естественного языка
  - –наиболее используемые типы формальных грамматик
  - –некоторые алгоритмы синтаксического разбора
  - –подходы к группировке

# Следующая лекция

- Статистические методы синтаксического анализа