

СИНТАКСИЧЕСКИЙ АНАЛИЗ: ПАРСЕРЫ

СОДЕРЖАНИЕ

1. Подходы к построению парсеров ЕЯ
 - базирующийся на грамматиках
 - базирующийся на статистике
2. Стратегии синтаксического анализа
 - для грамматик составляющих
 - для грамматик зависимостей
 - *предсинтаксический анализ*
3. Современные синтаксические парсеры
 - *StanfordParser, MaltParser* (для английского)
 - *ЭТАП, Дуалинг(АОТ), Comprero* (для русского)

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

ПОДХОДЫ К ПОСТРОЕНИЮ ПАРСЕРОВ

- *Парсер* - синтаксический анализатор
на входе: предложение текста
(результат морфологического анализа
словоформ)
на выходе: *синтаксическое дерево* предложения
- Подход, базирующийся на грамматических
правилах
 - Грамматика составляющих
 - Грамматика зависимостей
- Подход, базирующийся на статистике:
статистические анализаторы
- Современная тенденция: гибридные парсеры

ГРАММАТИЧЕСКИЕ МЕТОДЫ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Грамматика может быть
 - встроена в парсер
 - записана явно и отделена от процедуры разбора
- Для английского языка очень часто –
контекстно-свободные грамматики

КС-ГРАММАТИКА (определение)

КС-грамматика - это $\langle V_T, V_N, P, S \rangle$

- $V = V_T \cup V_N$ - множество терминальных и нетерминальных символов
- $P = \{A \rightarrow \beta : \beta \in V^*, A \in V_N\}$ - множество продукций
- S – начальный символ

Результат синтаксического анализа – это дерево:

- В корне дерева символ S
- В листьях символ из V_T
- В других узлах - символ из V

ПРИМЕР КС-ГРАММАТИКИ

$S \Rightarrow NP VP$

$N1 \Rightarrow Adj N1$

$N1 \Rightarrow N$

$N1 \Rightarrow N N$

$N1 \Rightarrow N N N$

$NP \Rightarrow N1$

$NP \Rightarrow Det N1$

$N1 \Rightarrow N1 PP$

$N1 \Rightarrow N1 ReCL$

$NP \Rightarrow Pron$

$NP \Rightarrow Name$

$VP \Rightarrow V$

$VP \Rightarrow V NP$

$VP \Rightarrow VP PP$

$PP \Rightarrow P NP$

$ReCL \Rightarrow WHN VP$

$ReCL \Rightarrow WHN S$

Нетерминалы соответствуют типам фраз и
обозначениям частей речи слов

Существенное развитие грамматических формализмов
для описания лингвистических явлений:

ГРАММАТИКИ ДЛЯ СА

- Грамматики составляющих:
 - Definite Clause Grammars (DCG)
 - Tree Adjoining Grammars (TAG)
 - Combinatory Categorical Grammars (CCG)
 - *Унификационные грамматики*:
 - *PATR* (формализм записи структур признаков)
 - *UTAG*
 - *HPDG* (Head-Driven Structure Grammar)
- Грамматики зависимостей:
 - Dependency Unification Grammars
 - Extensible Dependency Grammars (XDG)
 - Link Grammars (LG)
 - ...

МЕТОДЫ ПОСТРОЕНИЯ ГРАММАТИК

- Экспертами-лингвистами (вручную)
 - Построенная грамматика будет корректной.
 - Невозможно описать вручную все аспекты языка.
- Автоматизированно на основе корпусов текстов, с синтаксической разметкой; корпуса создаются вручную: *Treebank*
 - Возможность создать детальную грамматику

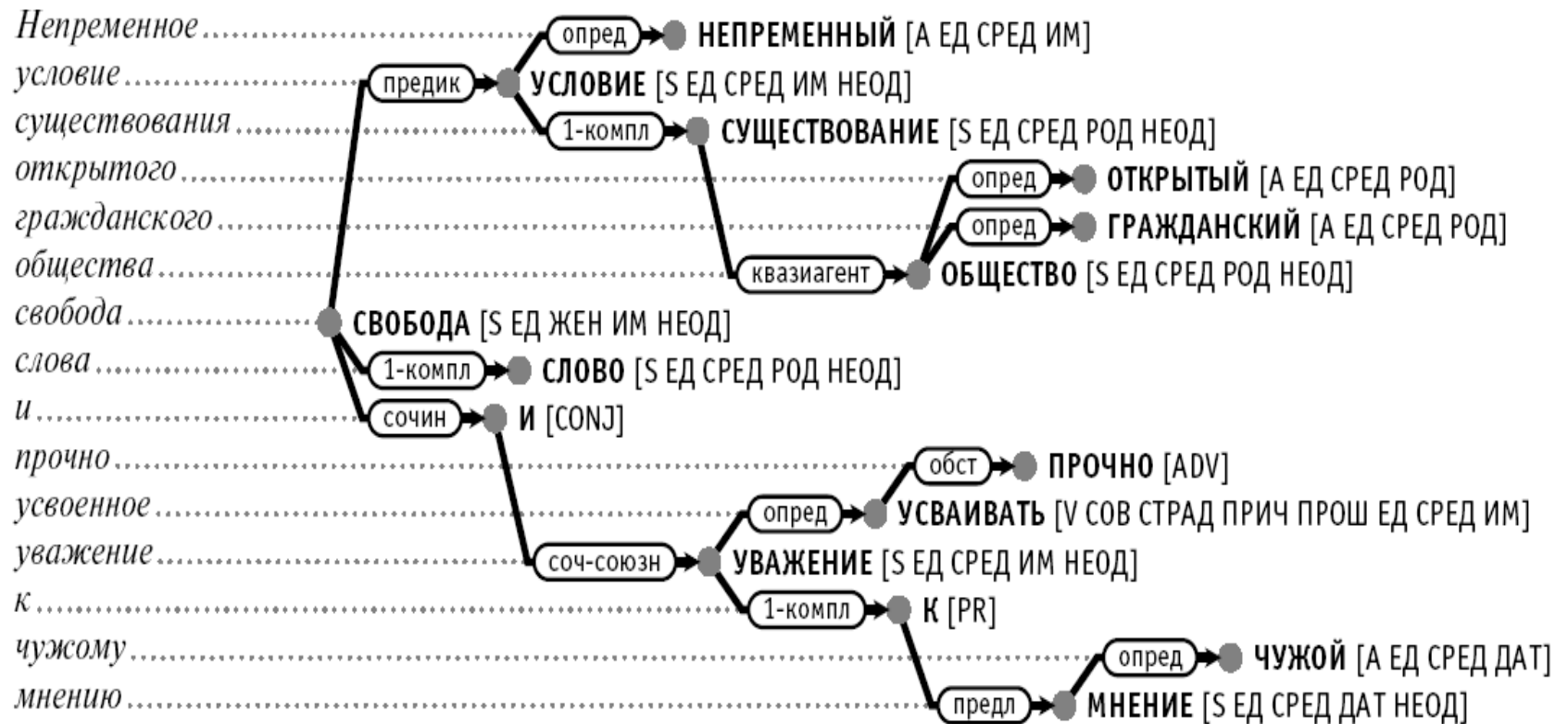
Однако:

 - Количество генерируемых грамматических правил слишком велико, что
 - Усугубляет проблему эффективности анализа и синтаксической омонимии.

ПРИМЕР СИНТАКСИЧЕСКОЙ РАЗМЕТКИ

- Национальный корпус русского языка:
ruscorpora.ru
- Синтаксическая разметка: *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, правильное дерево
зависимостей

Пример размеченного дерева из СинТагРус



СТАТИСТИЧЕСКИЕ АНАЛИЗАТОРЫ

- Основа – корпуса с синтаксической разметкой, *Treebank*
 - Для английского: *Penn Treebank*
 - Для чешского
 - Для русского (Национальный корпус РЯ)
- Оценка качества синтаксического анализа:
 - *S* – доля предложений с полностью правильным разбором
 - *W* – доля правильных главных слов для каждого слова
 - Правильный корень предложения
- Качество статистических парсеров:
 - Английский язык: *S* – 45%, *W* - 90-92%,
 - Чешский язык: *S* – 36%, *W* – 85%

СТРАТЕГИИ АНАЛИЗА ДЛЯ ГРАММАТИК ОСТАВЛЯЮЩИХ

- Алгоритмы синтаксического разбора на основе КС-грамматик:
 - Нисходящие (*top-down*)
 - Восходящие (*bottom-up*)
 - В общем случае – недетерминированный разбор с возвратами, экспоненциальная сложность
- Некоторые базовые алгоритмы
 - Алгоритм рекурсивного спуска (реализуется в виде рекурсивных функций, построенных на основе грамматики)
 - Алгоритм Кока-Янгера-Касами (СҮК)
 - Алгоритм Эрли

Пример: КС-грамматика

Grammar

$S \rightarrow NP VP$

$S \rightarrow Aux NP VP$

$S \rightarrow VP$

$NP \rightarrow Pronoun$

$NP \rightarrow Proper-Noun$

$NP \rightarrow Det Nominal$

$Nominal \rightarrow Noun$

$Nominal \rightarrow Nominal Noun$

$Nominal \rightarrow Nominal PP$

$VP \rightarrow Verb$

$VP \rightarrow Verb NP$

$VP \rightarrow VP PP$

$PP \rightarrow Prep NP$

Lexicon

$Det \rightarrow the \mid a \mid that \mid this$

$Noun \rightarrow book \mid flight \mid meal \mid money$

$Verb \rightarrow book \mid include \mid prefer$

$Pronoun \rightarrow I \mid he \mid she \mid me$

$Proper-Noun \rightarrow Houston \mid NWA$

$Aux \rightarrow does$

$Prep \rightarrow from \mid to \mid on \mid near \mid through$

Анализ снизу-вверх

book that flight

Анализ снизу-вверх

Noun

|
book

that

flight

Анализ снизу-вверх

Nominal

|

Noun

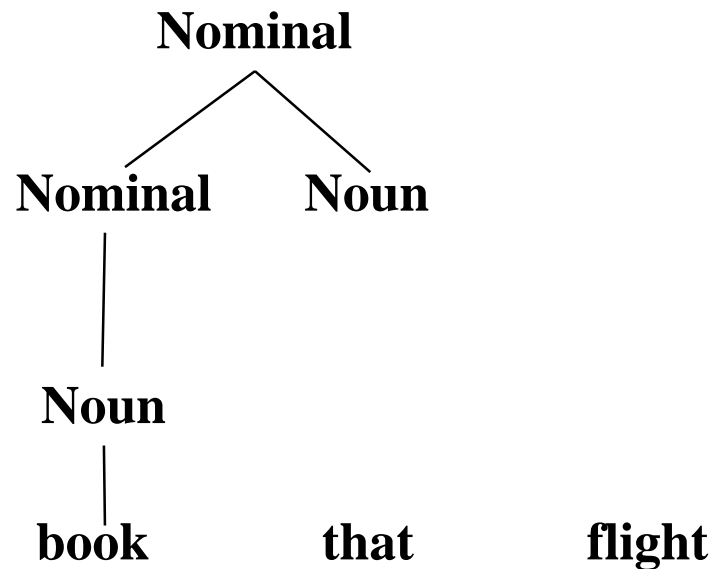
|

book

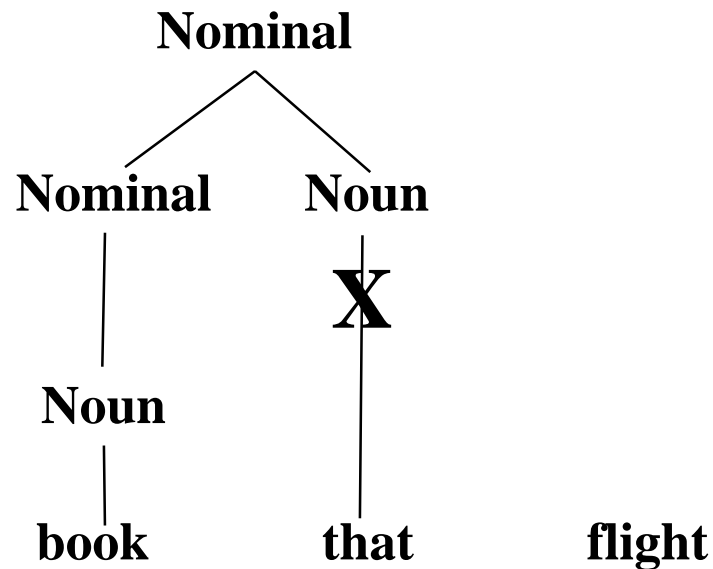
that

flight

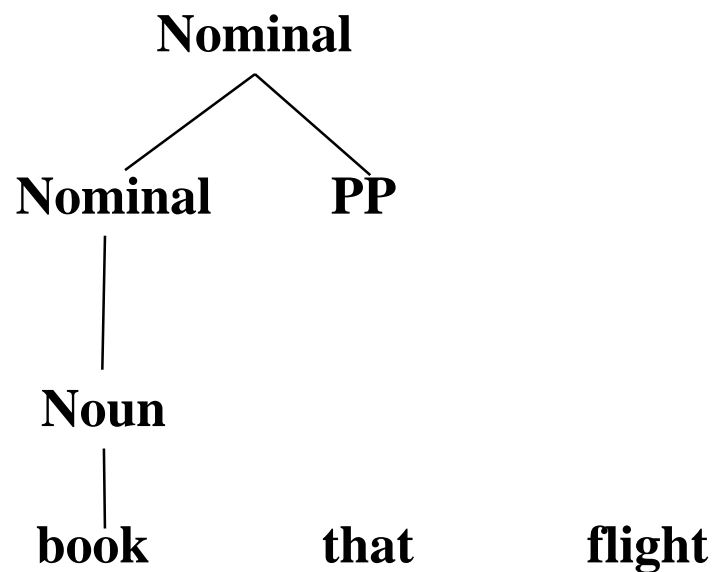
Анализ снизу-вверх



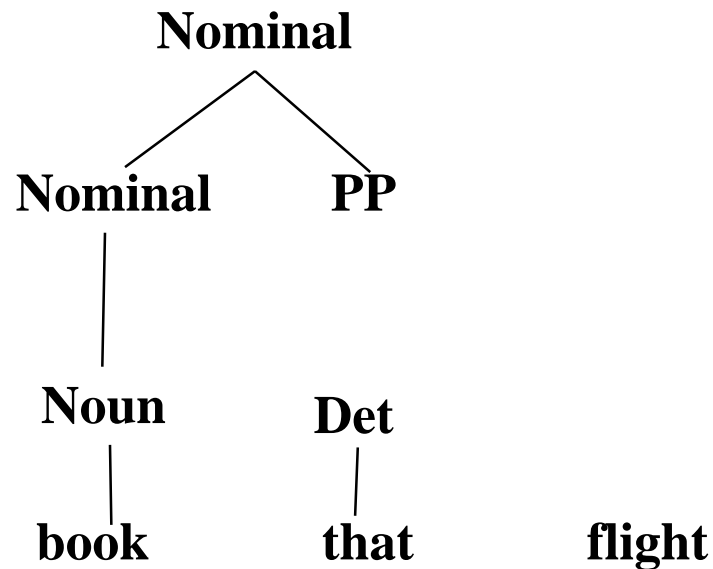
Анализ снизу-вверх



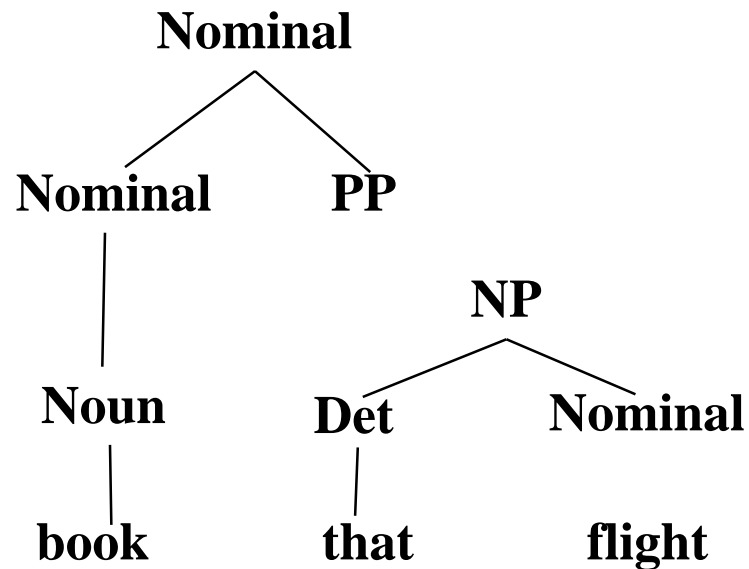
Анализ снизу-вверх



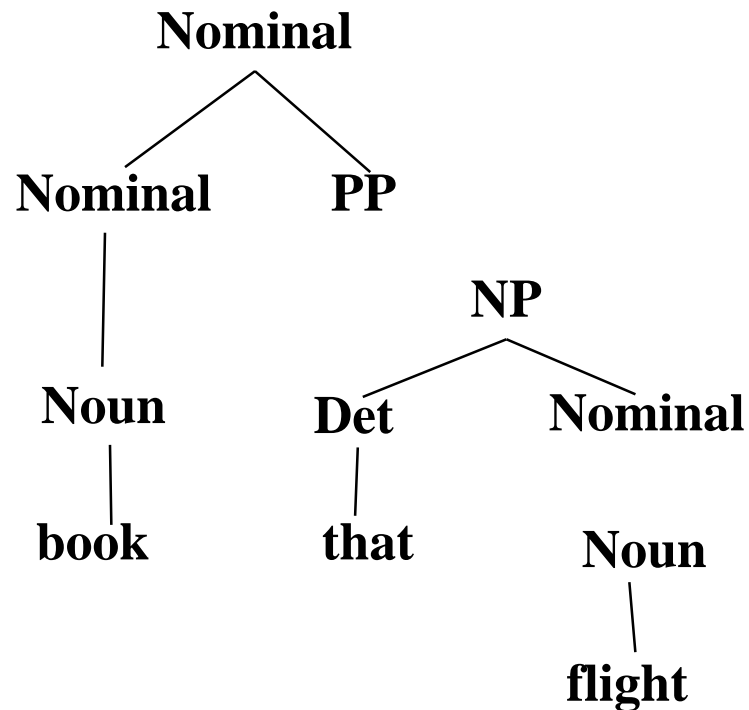
Анализ снизу-вверх



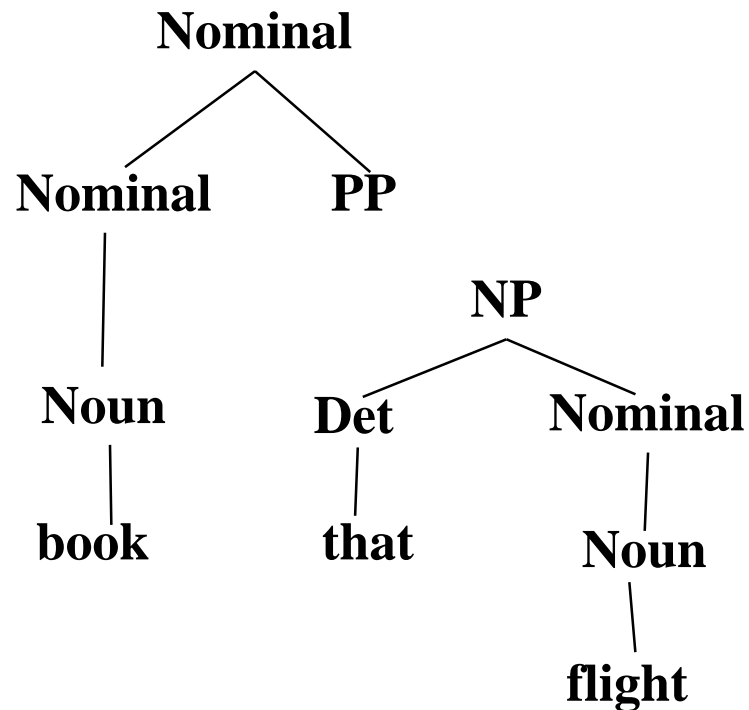
Анализ снизу-вверх



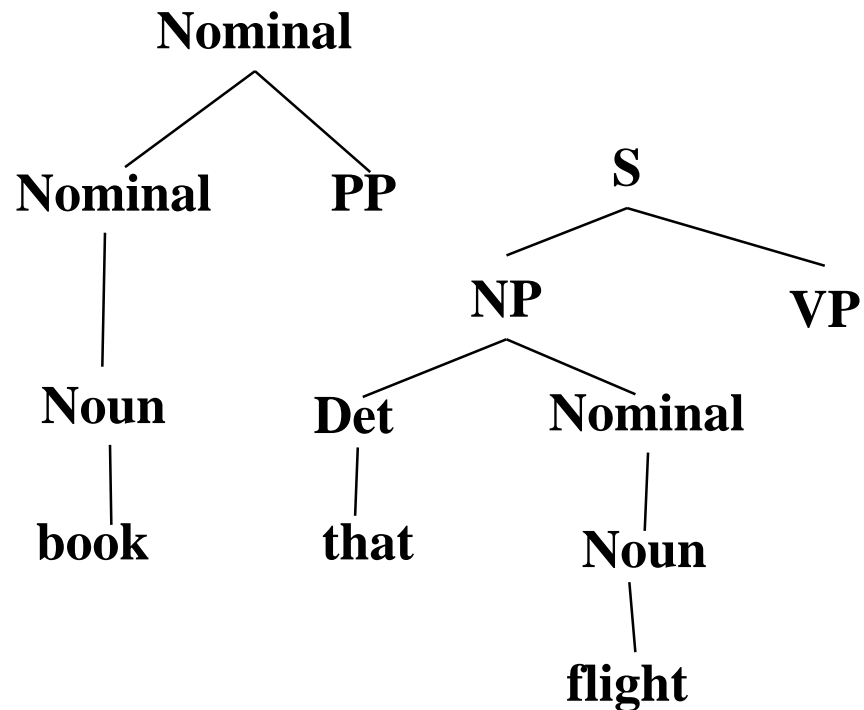
Анализ снизу-вверх



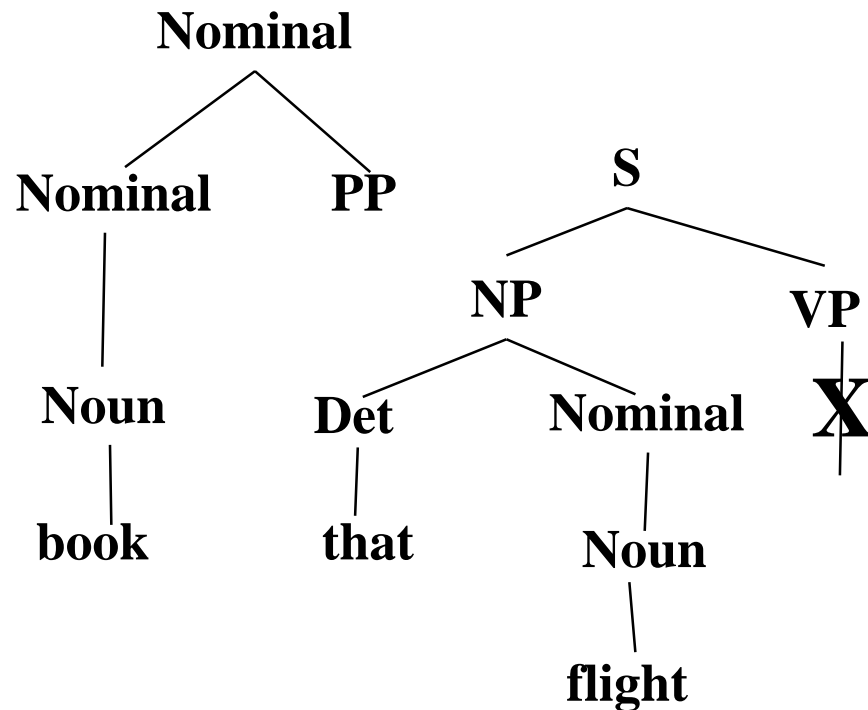
Анализ снизу-вверх



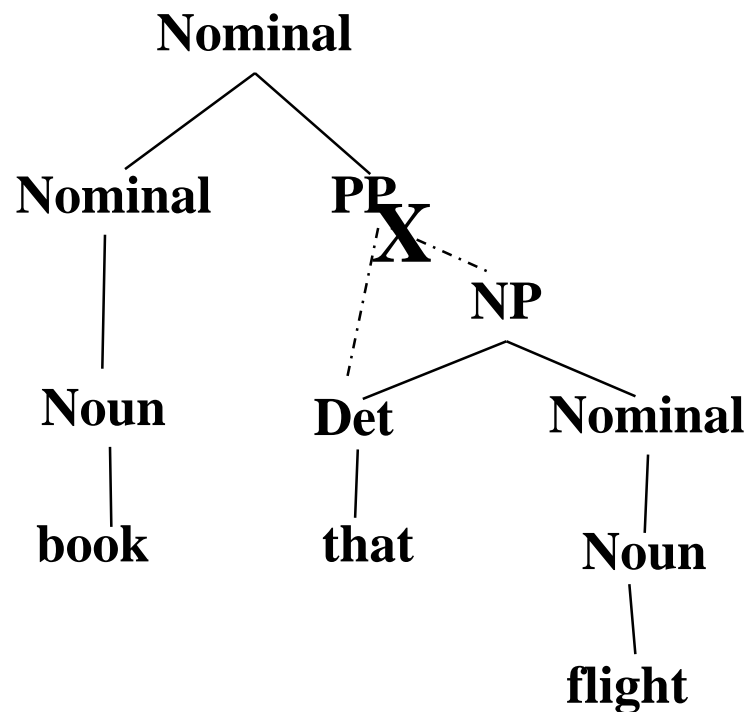
Анализ снизу-вверх



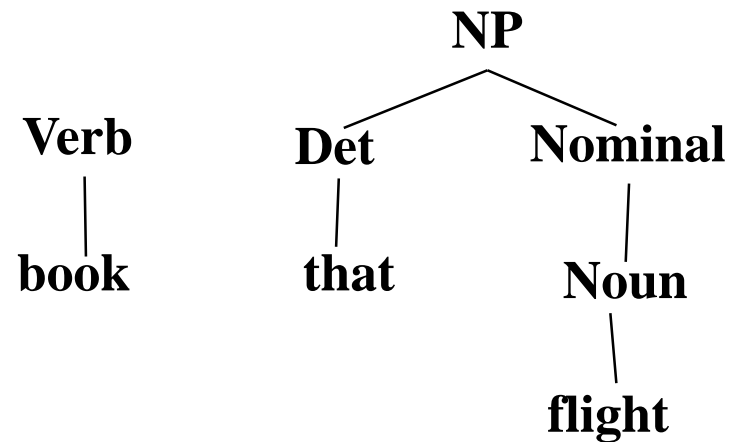
Анализ снизу-вверх



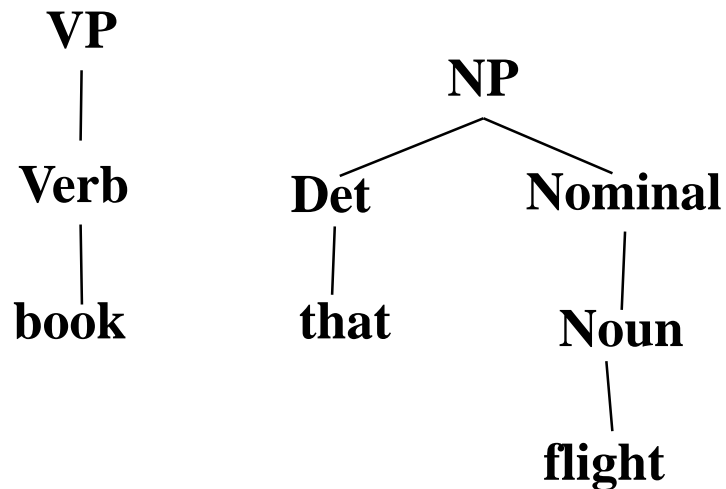
Bottom Up Parsing



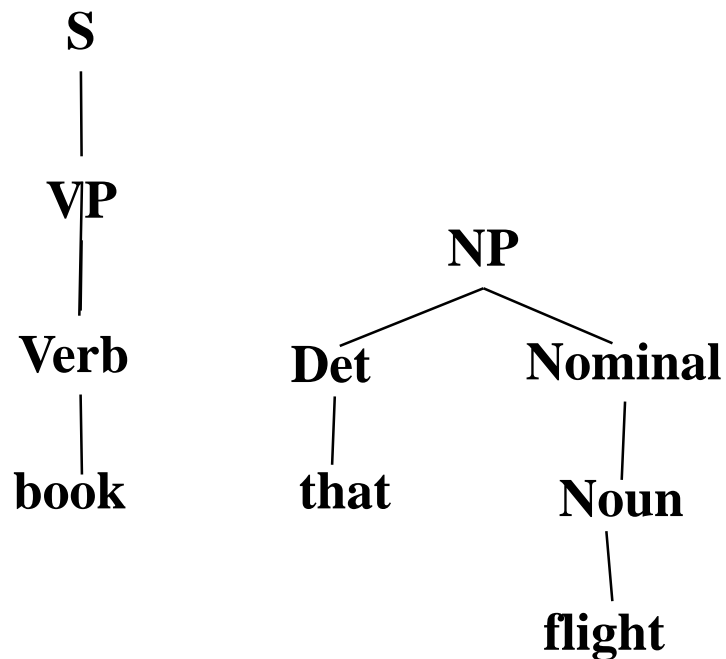
Bottom Up Parsing



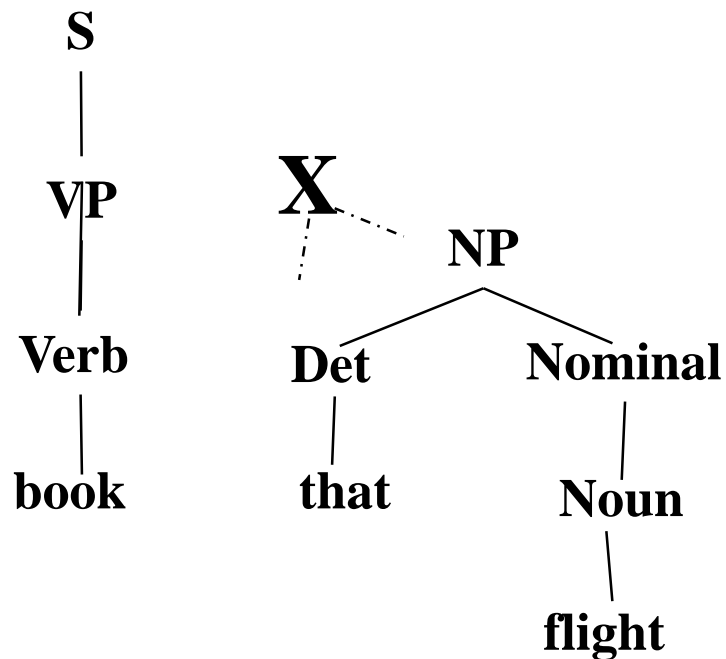
Анализ снизу-вверх



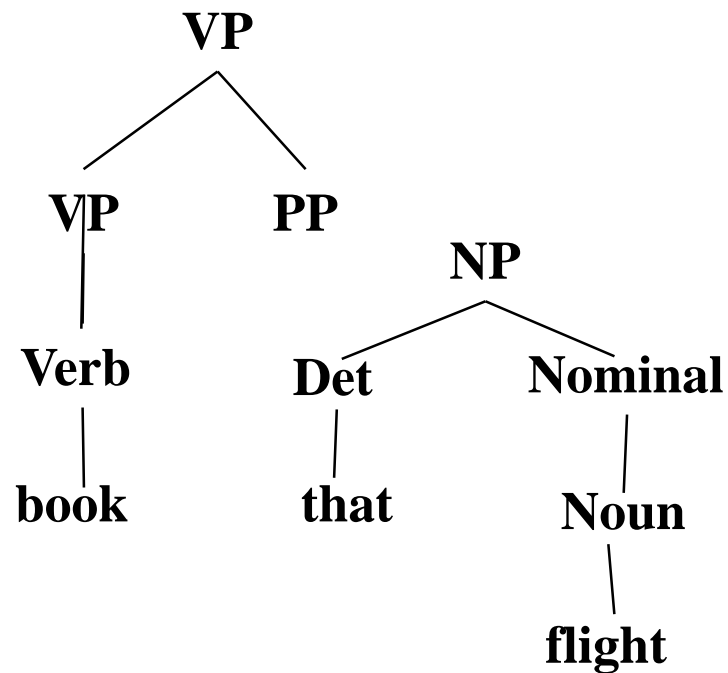
Анализ снизу-вверх



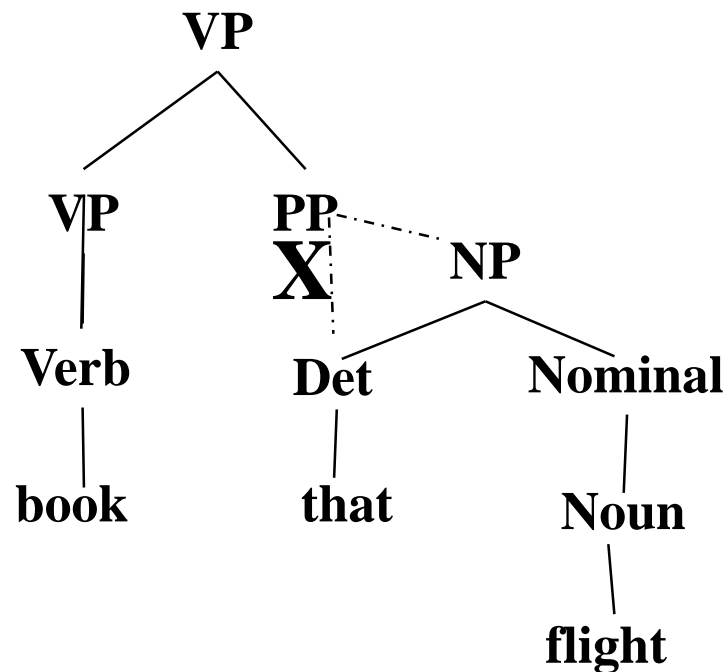
Анализ снизу-вверх



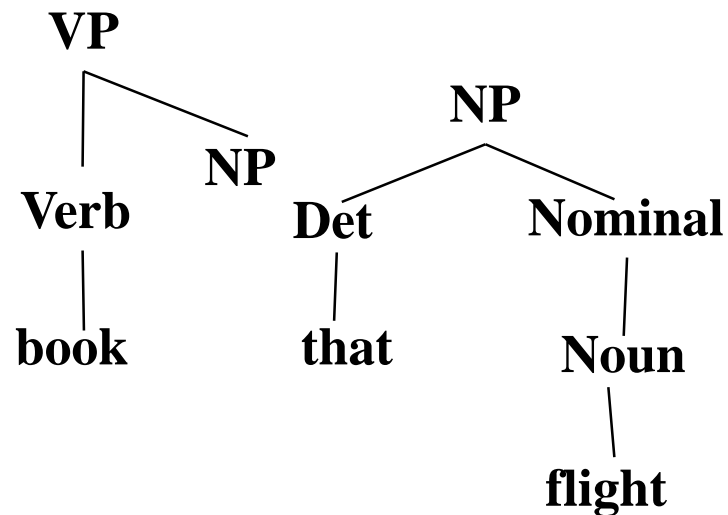
Анализ снизу-вверх



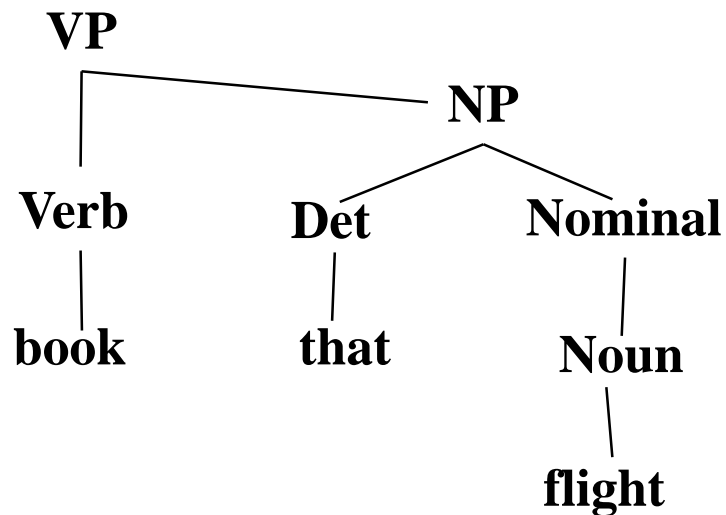
Анализ снизу-вверх



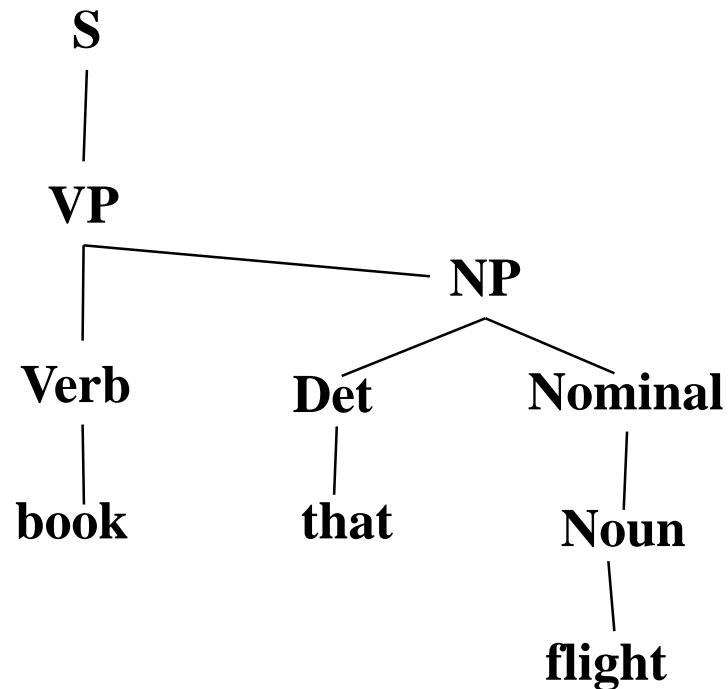
Анализ снизу-вверх



Анализ снизу-вверх



Анализ снизу-вверх



АЛГОРИТМЫ РАЗБОРА ПО КС-ГРАММАТИКАМ (динамическое программирование)

- Алгоритм Эрли
 - Разбор сверху-вниз
 - Предлагает деревья, которые не соответствуют словам
- Алгоритм Кока-Янгера-Касами (СҮК)
 - КС-грамматики в нормальной форме Хомского: правила вида $A \rightarrow BC$, $A \rightarrow \gamma$
 - Разбор снизу-вверх \Rightarrow деревья соответствуют словам
 - Возможны глобально бессмысленные деревья
 - Полиномиальная сложность: $O(|G| \times n \times n)$,
 n – длина предложения, $|G|$ – мощность грамматики

Пример грамматики в нормальной форме (Chomsky Normal Form)

Исходная грамматика

$S \rightarrow NP VP$

$S \rightarrow Aux NP VP$

$S \rightarrow VP$

$NP \rightarrow Pronoun$

$NP \rightarrow Proper-Noun$

$NP \rightarrow Det Nominal$

$Nominal \rightarrow Noun$

$Nominal \rightarrow Nominal Noun$

$Nominal \rightarrow Nominal PP$

$VP \rightarrow Verb$

$VP \rightarrow Verb NP$

$VP \rightarrow VP PP$

$PP \rightarrow Prep NP$

Chomsky Normal Form

$S \rightarrow NP VP$

$S \rightarrow X1 VP$

$X1 \rightarrow Aux NP$

$S \rightarrow book \mid include \mid prefer$

$S \rightarrow Verb NP$

$S \rightarrow VP PP$

$NP \rightarrow I \mid he \mid she \mid me$

$NP \rightarrow Houston \mid NWA$

$NP \rightarrow Det Nominal$

$Nominal \rightarrow book \mid flight \mid meal \mid money$

$Nominal \rightarrow Nominal Noun$

$Nominal \rightarrow Nominal PP$

$VP \rightarrow book \mid include \mid prefer$

$VP \rightarrow Verb NP$

$VP \rightarrow VP PP$

$PP \rightarrow Prep NP$

CKY Parser

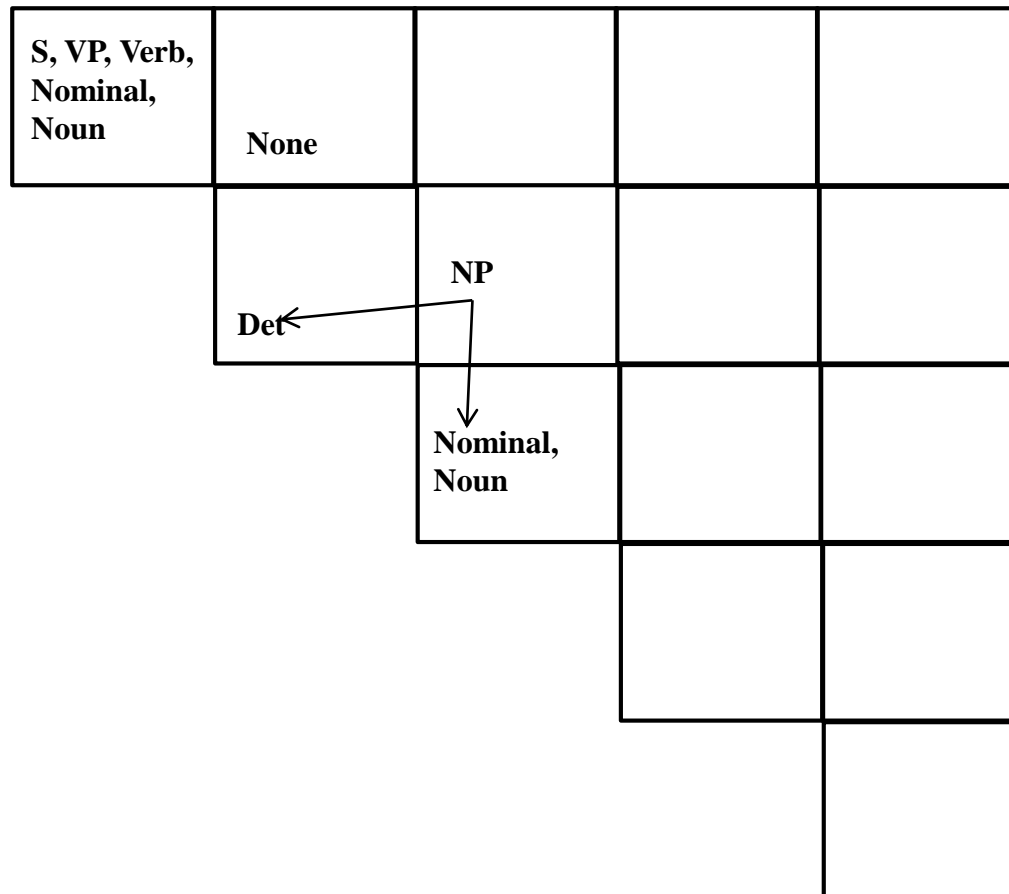
	Book	the	flight	through	Houston
	j= 1	2	3	4	5
i= 0					
1					
2					
3					
4					

Cell[i,j]
contains all
constituents
(non-terminals)
covering words
 $i + 1$ through j

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None			
		NP		
	Det			
		Nominal, Noun		



CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	VP		
	Det	NP		
		Nominal, Noun		

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP		
	Det	NP		
		Nominal, Noun		

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP		
	Det	NP		
		Nominal, Noun		

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP	None	
	Det	NP	None	
		Nominal, Noun	None	
			Prep	

CKY Parser

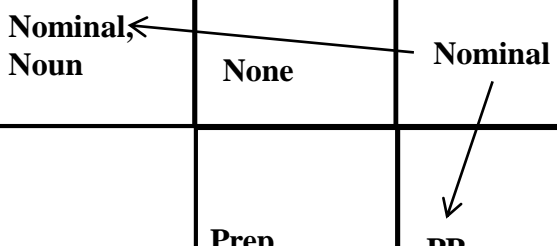
Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP	None	
	Det	NP	None	
		Nominal, Noun	None	
			Prep ←	PP
				↓ NP ProperNoun

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP	None	
	Det	NP	None	
		Nominal, Noun	None	Nominal
			Prep	PP
				NP ProperNoun



CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP	None	
		NP	None	NP
	Det ←			↓ Nominal
		Nominal, Noun	None	
			Prep	PP
				NP ProperNoun

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP	None	VP
	Det	NP	None	NP
		Nominal, Noun	None	Nominal
			Prep	PP
				NP ProperNoun

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP	None	S VP
	Det	NP	None	NP
		Nominal, Noun	None	Nominal
			Prep	PP
				NP ProperNoun

CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP ←	None	VP \$ VP
	Det	NP	None	NP
		Nominal, Noun	None	Nominal
			Prep	PP
				NP ProperNoun

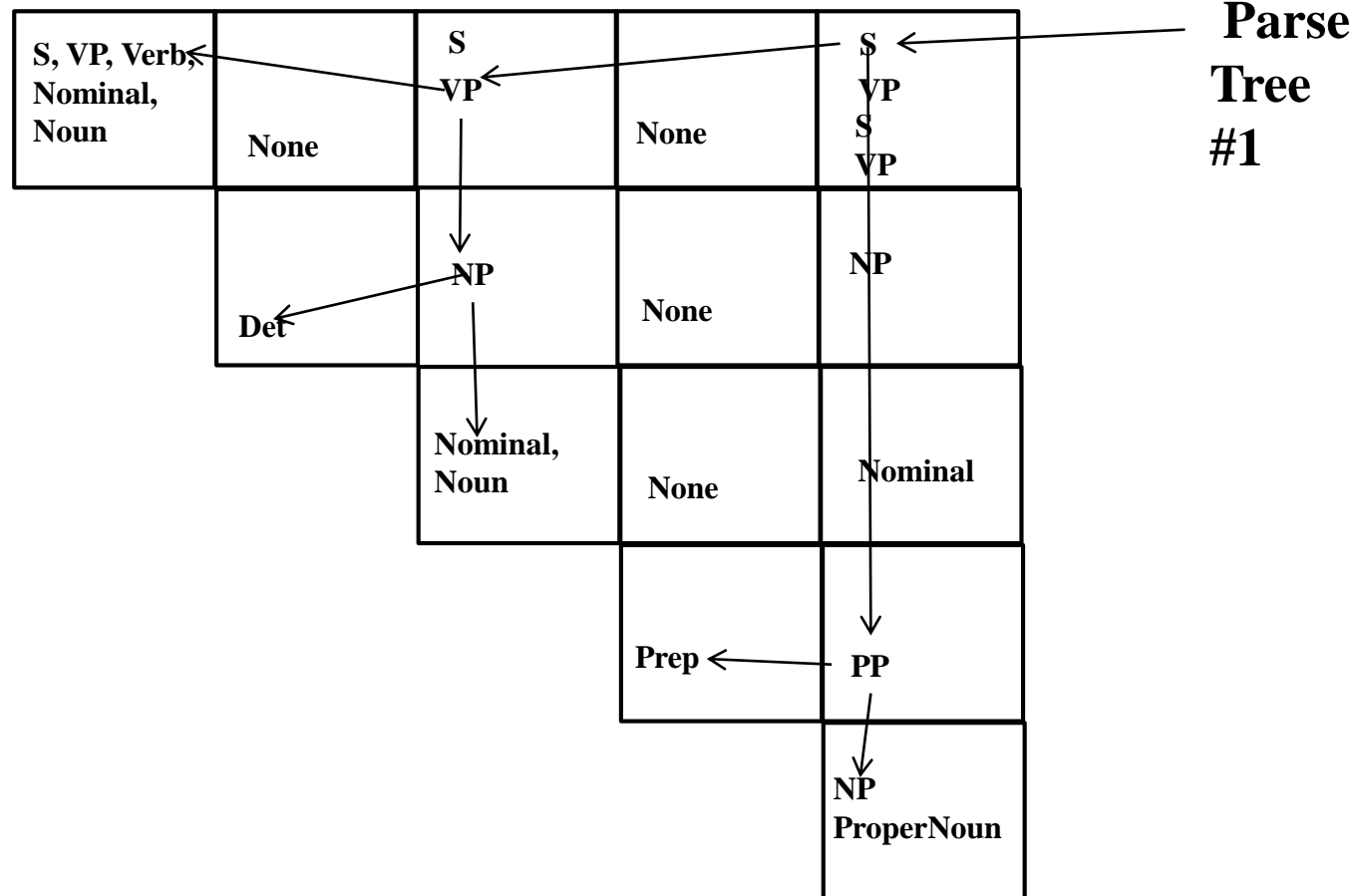
CKY Parser

Book the flight through Houston

S, VP, Verb, Nominal, Noun	None	S VP ←	None	S VP S VP
	Det	NP	None	NP
		Nominal, Noun	None	Nominal
			Prep	↓ PP
				NP ProperNoun

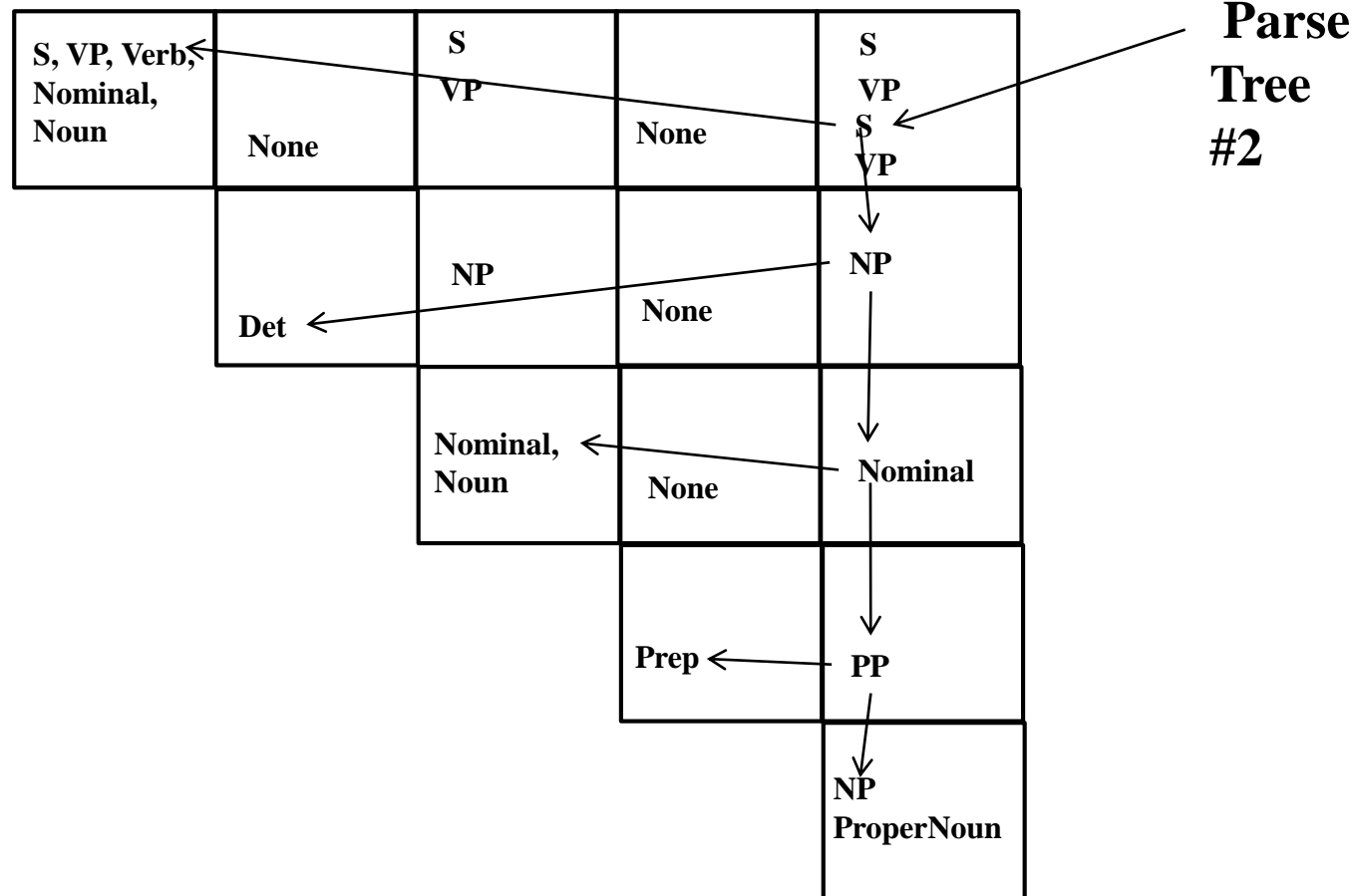
CKY Parser

Book the flight through Houston



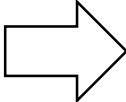
CKY Parser

Book the flight through Houston



СПОСОБ РАЗРЕШЕНИЯ СИНТАКСИЧЕСКИЙ ОМОНИМИИ

- Омонимия: несколько синтаксических деревьев
- На основе корпуса подсчитать вероятность каждого правила грамматики

$$\begin{array}{ccc} P = \{ \dots & & P = \{ \dots \\ & NP \rightarrow DT \ NN & 0,3 \\ & NP \rightarrow DT \ ADJ \ NN & 0,6 \\ & NP \rightarrow NN \ NN & 0,1 \\ & \dots \} & \end{array}$$


- Вероятность дерева разбора определяется перемножением вероятностей правил, примененных при его построении
- Выбирается наиболее вероятное дерево

СОВРЕМЕННЫЕ МОДЕЛИ СА

Учитывают ограничения (*constraints*),
накладываемые на соединение языковых единиц,
существующие в большинстве ЕЯ

⇒ развитие моделей и грамматик.

- **Согласование** (*agreement*) слов языка, например:
(нет) большого самолета
 - Широко представлено во флективных языках – согласование морфологических параметров слов: рода, падежа, числа и др.
 - В западной лингвистике возникли понятия: *feature structure, unification of feature structures* (привлечена процедура логической унификации для согласования свойств/признаков слов)
- **Валентность** как общая сочетательная способность слов и других языковых единиц
(сопоставимо с понятием предиката в логике)

СОВРЕМЕННЫЕ МОДЕЛИ СА: ВАЛЕНТНОСТЬ

- *Валентность* – способность слова присоединять другие единицы определенным синтаксическим способом
- *Слова-предикаты* описывают ситуации и действия:
 - глаголы и глагольные формы: *идти, приходящий*
 - отглагольные существительные: *преобразование*
 - краткие прилагательные: *рад, должен*
 - предлоги: *к (морю)*
- Слова-предикаты имеют места для заполнения - *валентности*, например:
 - Подарить: *кто? (1) что? (2) кому? (3)*
 - Рубить: *кто? (1) что? (2) чем? (3)*
- *Актант* – заполнитель валентности:
слово, словосочетание, фраза
- Валентности отличаются по степени обязательности

ВАЛЕНТНОСТИ И МОДЕЛИ УПРАВЛЕНИЯ

Валентности по разному описываются
в рамках двух подходов к СА:

- Структуры/деревья зависимостей:
 - **Модель управления глагола** (слова-предиката) – набор его валентностей.
 - Обычно определены синтаксические ограничения на актанты, например: падеж актантов.
 - Важна информация о взаимном расположении актантов.
- Структуры/деревья составляющих (западная КЛ)
 - **Subcategorization frames** *John gives him a book*
John gives a book to him
 - **Субкатегоризация** - выделение подкласса синтаксической (фразовой) категории со специфическими свойствами
- Слово может иметь несколько моделей управления
- Валентности: синтаксические и семантические

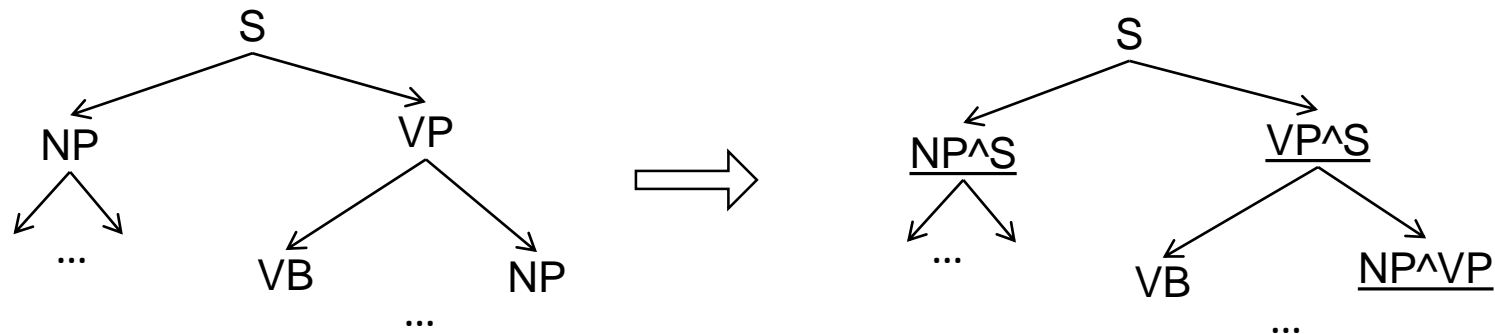
Синтаксическая модель управления (таблица)

- Наказывать

1=A	2=B	3=C	4=D
S _{ИМ}	S _{ВИН}	за S _{ВИН}	S _{ТВ}

ПУТИ РЕШЕНИЯ ПРОБЛЕМЫ НЕОБНОЗНАЧНОСТИ ГРАММАТИКИ

Добавление контекстной и родительской информации у составляющих (нетерминалов)



Однако:

- Увеличивается общее количество продукций.
- Уменьшается количество продукции для конкретной составляющей.

СТРАТЕГИИ АНАЛИЗА ДЛЯ ГРАММАТИК ЗАВИСИМОСТЕЙ

- Деревья зависимостей:
 - Каждое слово-узел зависит от одного слова, кроме корневого узла
 - Зависимости не образуют циклов
 - Выполняется свойство проективности
- Метод фильтров:
 - Порождаются всевозможные синтаксические связи слов
 - Отбрасываются ошибочные и избыточные связи путем применения фильтров, например, правил согласования
 - Фильтры задают условия, описывающие правильно построенные деревья
- На практике для эффективности применяются:
 - предсинтаксический анализ:
синтаксическая сегментация
 - установление высоковероятных локальных связей

ЛОКАЛЬНЫЕ ВЫСОКОВЕРОЯТНЫЕ СИНТАКСИЧЕСКИЕ СВЯЗИ ДЛЯ 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*
- *Num* и *Num*: *тридцать* ← *три*

ПРАВИЛА УСТАНОВЛЕНИЯ ЛОКАЛЬНЫХ СВЯЗЕЙ

Правило установления 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

ПРАВИЛО УСТАНОВЛЕНИЯ ЛОКАЛЬНОЙ СВЯЗИ *A* И *N*

Правило установления локальной связи
прилагательное-существительное:

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 $A \leftarrow N$

Локальные связи - высоковероятные

ПРЕДСИНТАКСИЧЕСКИЙ АНАЛИЗ

Синтаксическая сегментация часто предшествует собственно синтаксическому анализу

- Понятие сегментации в КЛ (*Segmentation*)
- Основные виды сегментации:
 - ❖ Сегментация нижнего уровня (*low-level*)
 - ❖ Сегментация высокого уровня (*high-level*)
- Сегментация нижнего уровня (уровень символов)
 - Выделение слов (псевдослов) в потоке знаков - *tokenization* (*графематический анализ*)
 - Разбиение текста на предложения
- Сегментация высокого уровня:
синтаксическая, композиционная

СЕГМЕНТАЦИЯ СИНТАКСИЧЕСКАЯ И КОМПОЗИЦИОННАЯ

- **Синтаксическая сегментация (*inter-sentence*) - *syntactic chunking*:**
 - выделение простых предложений в составе сложных для проведения их независимого синтаксического анализа
 - выделение локальных синтаксических групп – именных, глагольных и др.
 - по сути – частичный синтаксический анализ, вычислительная сложность – $O(n)$
- **Композиционный анализ (*intra-sentence*):**
 - выделение композиционных элементов:
 - абзацы и рубрики
 - заголовки разделов и подразделов
 - эпиграфы, сноски, примечания
 - установление иерархии абзацев, рубрик, предложений

Конкретные синтаксические анализаторы

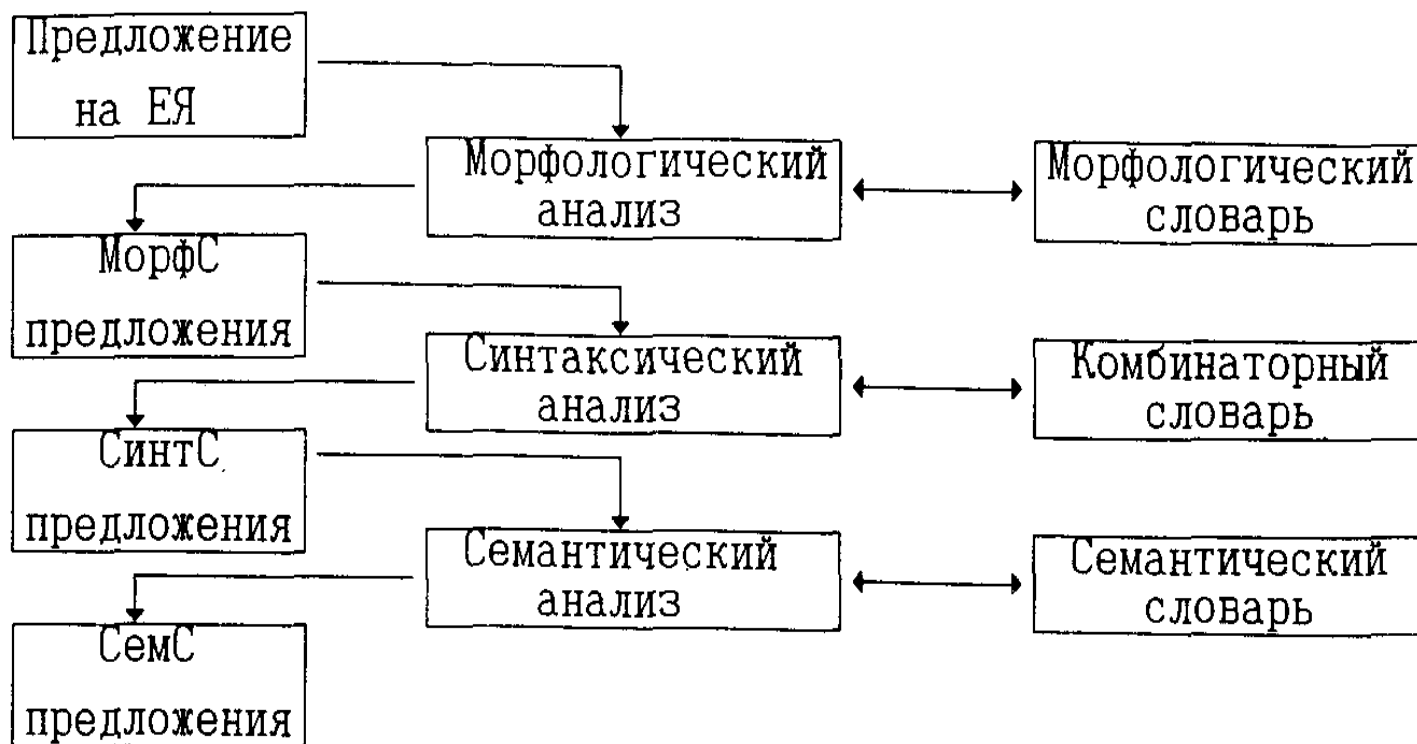
ПАРСЕРЫ ДЛЯ АНГЛИЙСКОГО ЯЗЫКА

- *Stanford Parser* — <http://nlp.stanford.edu/software/lex-parser.html>
 - Грамматика составляющих
 - Нисходящий синтаксический анализ
 - Использование вероятностей и контекстной информации о правилах грамматики
 - Включает средства преобразования дерева составляющих в дерево зависимости
- *MaltParser* — <http://maltparser.org/>
 - Статистический анализатор
 - Грамматика зависимостей
 - Метод опорных векторов для обучения модели анализа
 - Несколько версий для разных ЕЯ

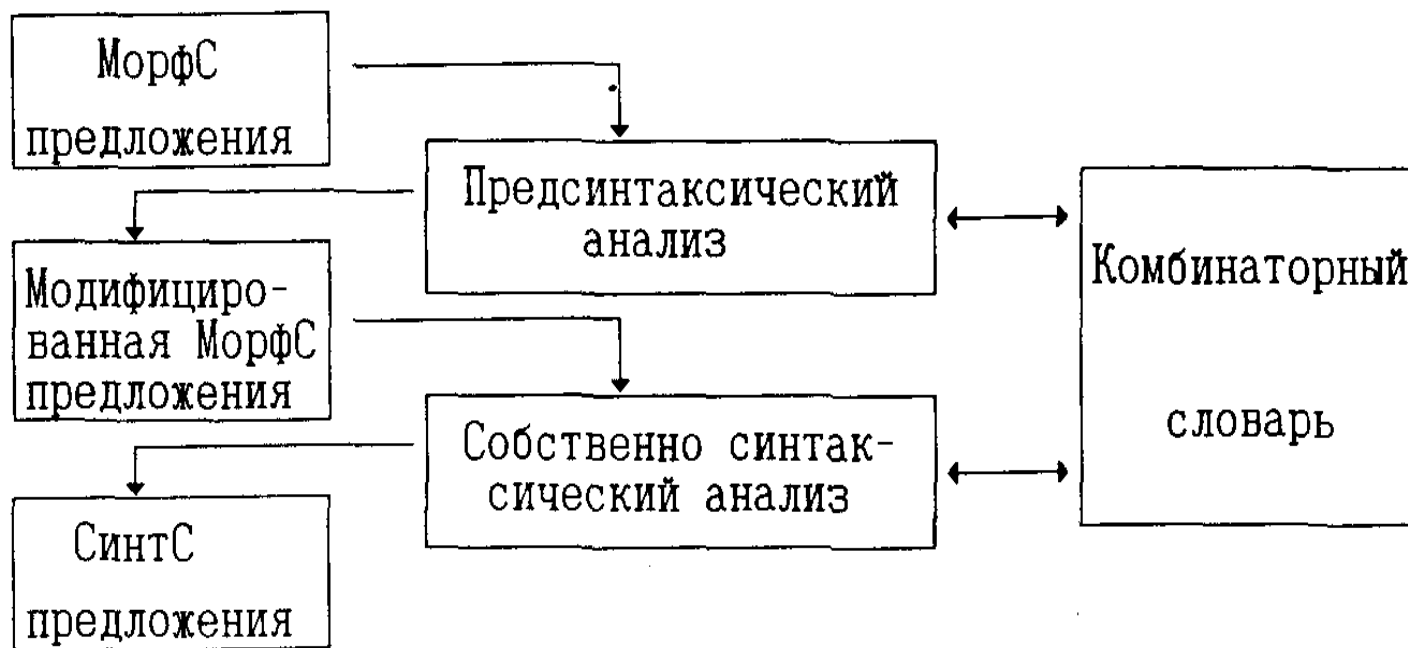
ПАРСЕРЫ ДЛЯ РУССКОГО ЯЗЫКА: СИСТЕМА ЭТАП

- Парсер системы машинного перевода *Этап-3*
<http://proling.iitp.ru/etap/>
(Институт Передачи Информации РАН)
- Лингвистическая теория (модель)
«Смысл \Leftrightarrow Текст»: И.А. Мельчук, А.Д. Апресян
- Множество разного вида лингвистических правил анализа
- *ТКС – толково-комбинаторный словарь*
уровень синтаксиса:
- *модели управления* слов-предикатов,
т.е. описание их синтаксических валентностей,
в частности: синтаксические ограничения на
актанты, например: падеж актантов
(*актант* – заполнитель валентности: слово,
словосочетание, фраза)

ЭТАП: ОБЩАЯ СХЕМА АНАЛИЗА



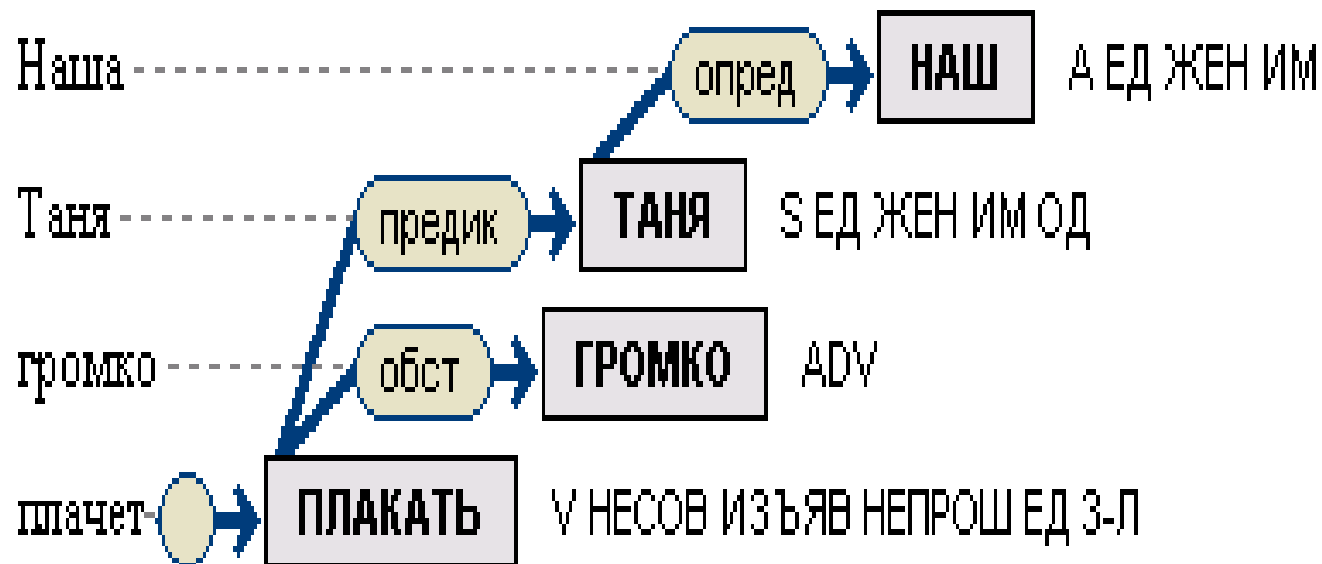
ЭТАП: СИНТАКСИЧЕСКИЙ АНАЛИЗ



ЭТАП: СИНТАКСИЧЕСКИЕ СТРУКТУРЫ И ОТНОШЕНИЯ

- Синтаксическая структура предложения – размеченное дерево зависимостей:
 - узлы дерева – слова предложения;
 - каждая дуга дерева помечена именем синтаксического отношения : *СинтО*
- Все СинтО являются бинарными и ориентированными
- Типы СинтО:
 - Актантные отношения
 - *Предикативное*: сказуемое → подлежащее
 - Атрибутивные отношения
 - *Определительное*:
существительное → прилагательное (причастие)
 - Сочинительные отношения
 - Служебные отношения

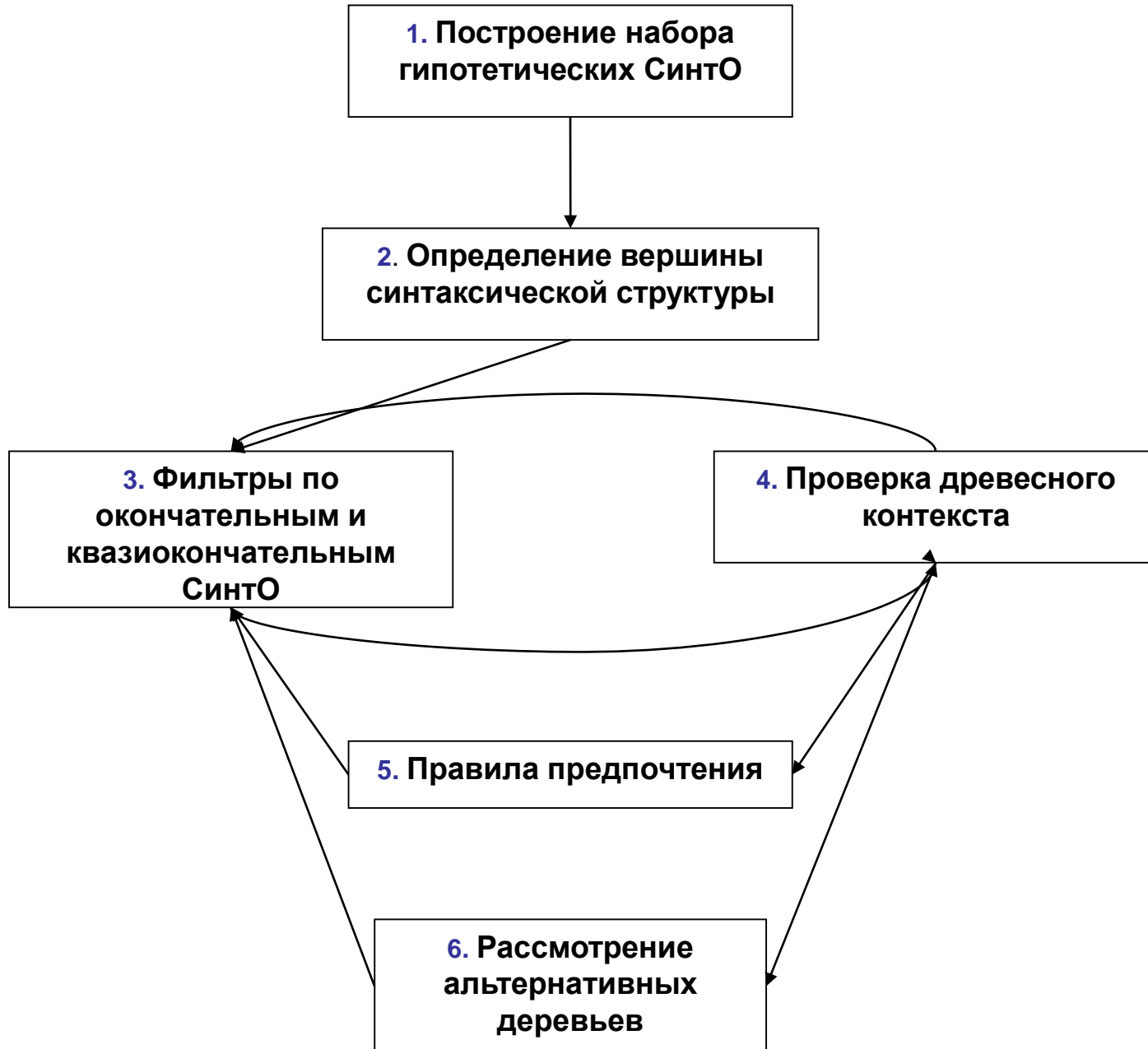
ЭТАП: ПРИМЕР СИНТАКСИЧЕСКОЙ СТРУКТУРЫ



ЭТАП: ПРЕДСИНТАКСИЧЕСКИЙ АНАЛИЗ

- Разрешение морфол. омонимии, например:
 - наречие/краткое прилагательное ср. рода: *было хорошо*.
Правило: если непосредственно слева или справа от слова (на расстоянии не более трех слов) есть несвязочный глагол (глагол, отличный от *быть, оказаться, становиться*), то наречие исключается.
- Установление конкретной синтаксической связи, например:
 - *ограничительное* СинтО между частицей *НЕ* и непосредственно следующим за ней глаголом в личной форме или инфинитиве: *не пишет, не писать*
- Определение вспомогательной характеристики:
 - выделение потенциального главного слова в предложении, если нет глагола:
Задача решена. ... Кто начальник торгового

ЭТАП: СХЕМА АНАЛИЗА



ЭТАП: ПОСТРОЕНИЕ СВЯЗЕЙ

- Построение гипотетических СинтО между словами предложения
 - на основе правил *синтагм* (= неразрывных синтаксических единств):
учет линейного контекста и морфопризнаков слов;
 - в результате – ориентированный граф гипотетических СинтО, число дуг (ветвей) графа обычно в 2-4 раза превосходит конечное число.
- Определение вершины (корня) синтаксической структуры
 - просмотр – в какие узлы не входит ни одна дуга;
 - если несколько возможных вершин - омонимам всех слов присваивается вес потенциальной возможности, что они являются вершиной.

ЭТАП: ФИЛЬТРАЦИЯ

- Фильтрация СинтО
 - если у некоторого слова (отличного от вершины) есть омонимы, не имеющие ни одного синтаксического хозяина, то эти омонимы стираются.
 - если в графе имеется узел, куда входит только одна дуга, то она помечается как окончательное СинтО.
 - если в графе имеется узел, куда входит несколько дуг, но все они исходят из одного и того же узла графа, то эти дуги помечаются как квазиокончательные (два слова связаны, но непонятно, какой именно связью)
- Уничтожаются лишние омонимы той пары слов, элементы которой связаны окончательной связью, кроме омонимов, образующих именно данную связь
- Проверка древесного контекста - установленных гипотетических и окончательных СинтО
 - в один узел входит не более одной связи
 - из узла выходит не более одной связи с данным именем
 - проективность синт. дерева
(в некоторых конструкциях допускается непроективность)³⁶

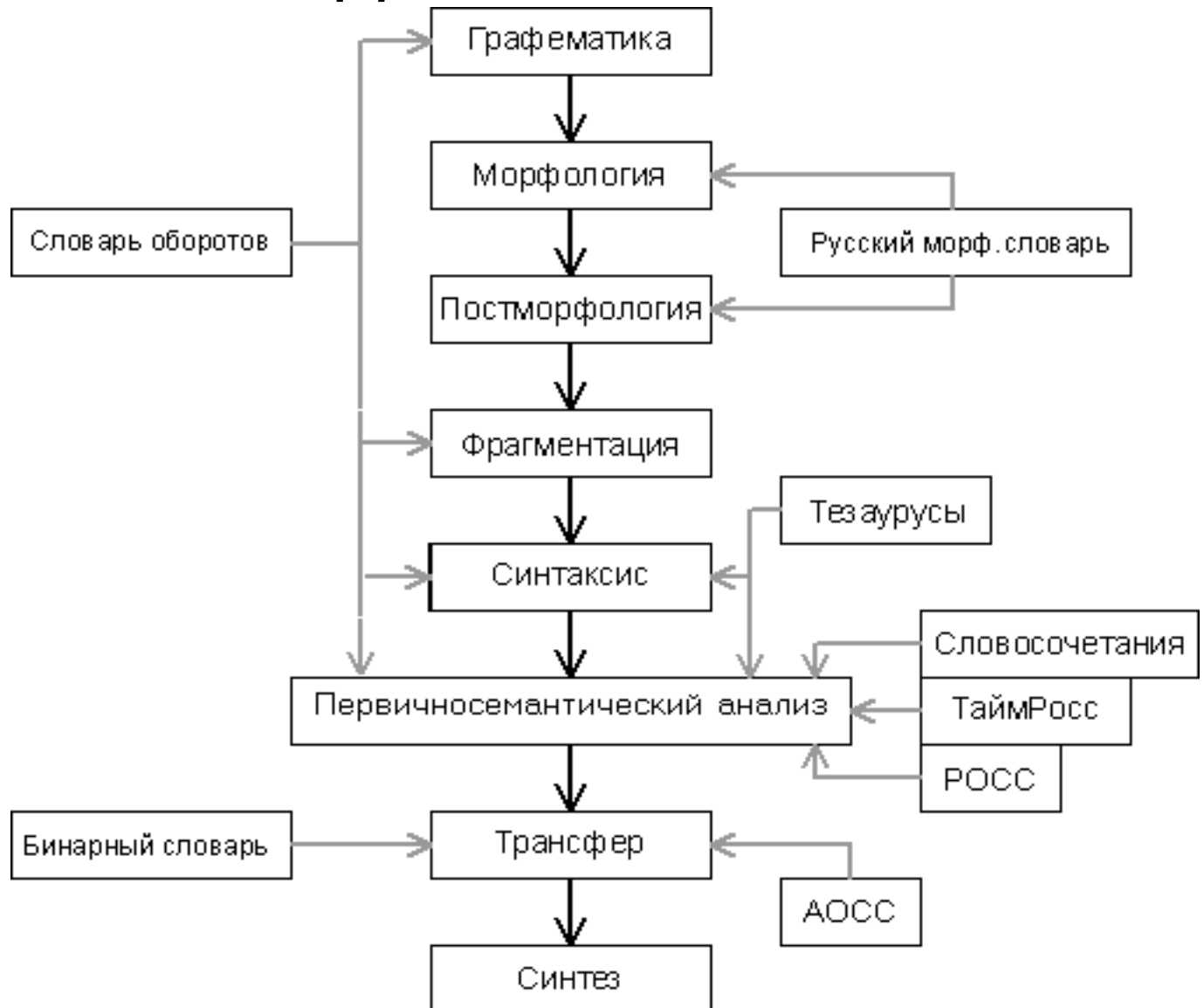
ЭТАП: ПРАВИЛА ПРЕДПОЧТЕНИЯ

- Фильтрация и проверка древесного контекста работают в цикле до тех пор, пока производятся изменения в графе (обычно 1-3 прохода)
 - одни связи предпочитаются, другие уничтожаются
- Рассмотрение альтернативных деревьев, если граф СинтО не является деревом – для выбора одного дерева разбора
 - Рассматриваются узлы, в которые входят более чем одна дуга: в искомое дерево оставляется только одна (связывающая наиболее близкие в предложении слова), остальные дуги стираются.

ПАРСЕРЫ ДЛЯ РУССКОГО ЯЗЫКА: ДИАЛИНГ / АОТ

- Синтаксический процессор Диалинг:
Л. Гершензон, Д. Панкратов, А. Сокирко, 1998-2001 гг.
- Проект АОТ лингвистического анализа русскоязычных текстов: aot.ru
 - включает работы по проекту Диалинг
 - программные модули с открытым кодом, лицензия *LGPL*
- Синтаксический анализ:
используется понятие синтаксической группы
(комбинированная, гибридная модель синтаксиса)
- Демонстрация анализа в режиме он-лайн:
<http://www.aot.ru/demo/synt.html>
- Модуль *SynAn* пакета Dialing

МОДУЛИ АОР



АОТ: ОСОБЕННОСТИ АНАЛИЗА

- Модуль графематики выполняет:
 - токенизацию
 - сегментацию на предложения
 - свертку некоторых словосочетаний
 - композиционный анализ текста (напр., выделение абзацев)
- Морфологический модуль: каждая словоформа представлена множеством морфологических ОМОНИМОВ
- Синтаксический анализ:
 - Взаимодействие модуля **фрагментации** (синтаксической сегментации) и собственно **синтаксиса** (построение синтаксических групп).
 - Не ставится цель получить полную синтаксическую структуру предложения, только формирование различных синтаксических групп слов.
 - Использование моделей управления слов-предикатов.
 - Синтаксические правила представлены процедурно.

АОТ: ФРАГМЕНТАЦИЯ

- Фрагментационный анализ – деление предложения на неразрывные синтаксические единства и установление частичной иерархии
 - ❖ главные и придаточные предложения (простые)
 - ❖ причастные и деепричастные обороты
- Границы фрагментов не должны пересекать синтаксических связей
- Примеры правил этого этапа:
 - Правила, уничтожающие омонимию
 - Если есть неомонимичное слово-предикат (глагол, причастие и др.), то во всех остальных словах данного фрагмента уничтожаются предикативные омонимы: *Мыла на кухне она не нашла*
 - Правила, устанавливающие иерархию
 - ...тот, кто этого не знает, не решит*
 - Правила объединения дистантных фрагментов

АОТ: СИНТАКСИЧЕСКИЕ ГРУППЫ (39 типов)

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

АОТ: НАЧАЛО СИНТАКСИЧЕСКОГО АНАЛИЗА

1. Первичная сегментация предложения по знакам пунктуации и сочинительным союзам с учетом простейших рядов однородных членов
2. Выделение аналитических форм глагола: *будет играть*
3. Выделение терминологических именных словосочетаний (на базе тезаурусов)
4. Обработка существующих и восстановление пропущенных тире в функции связи
5. Построение множества морфологических интерпретаций (МИ) для выделенных сегментов
6. Выявление для каждой МИ выделенных сегментов простых групп вида ПРИЛ-СУЩ, КОЛИЧ, ПГ

АОТ: ПРОДОЛЖЕНИЕ

СИНТАКСИЧЕСКОГО АНАЛИЗА

7. Объединение сочиненных сегментов и построение сочиненных синтаксических групп (именных, глагольных) внутри сегментов для каждой МИ
8. Установление зависимости между соседними сегментами
(выявление отношений подчинения, вложения)
9. Построение синтаксических групп, включающих вложенные сегменты
10. Объединение разрывных сегментов предложения
11. Выявление зависимостей и построение синтаксических групп с использованием всех синтаксических правил
12. Ранжирование результатов анализа для разных МИ, выбор наилучшего варианта разбора

ПАРСЕРЫ ДЛЯ РУССКОГО ЯЗЫКА:

Compreno

- Разрабатывается в АВВУУ более 15 лет
- 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 непроективные связи.

Compreno:

ПРИМЕР СИНТАКСИЧ. АНАЛИЗА

