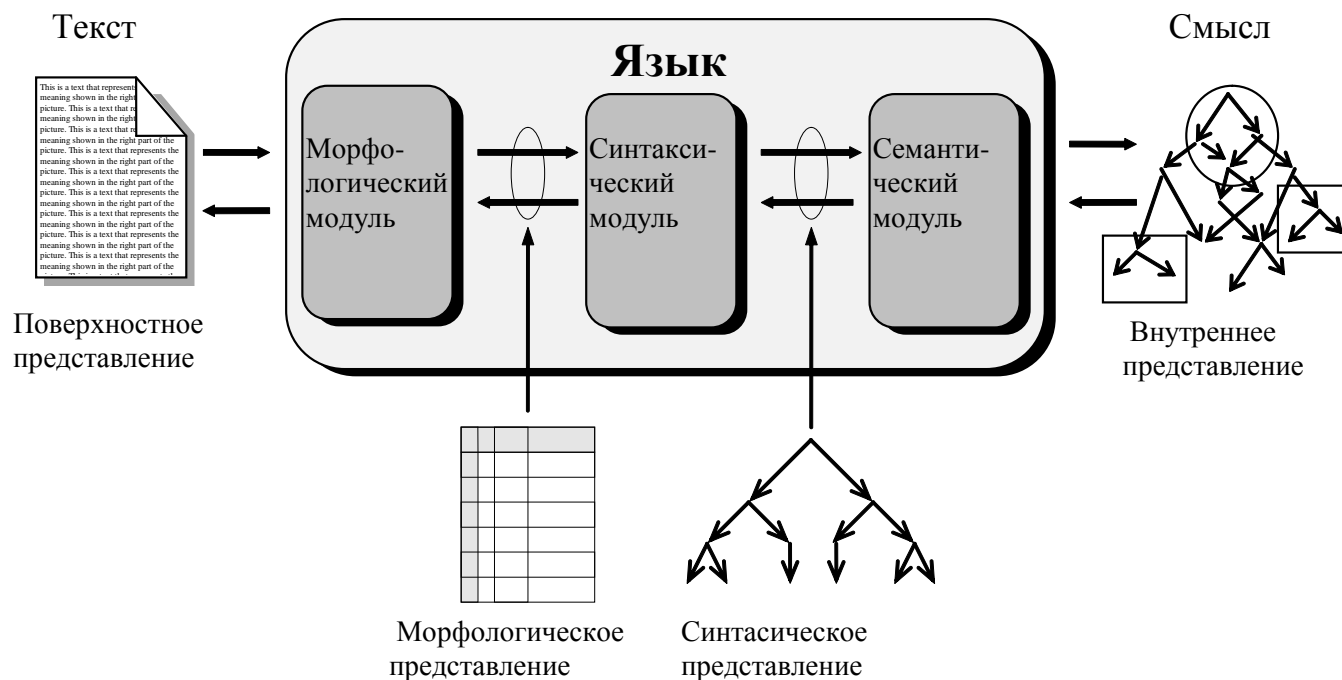


# Синтаксический анализ ЕЯ: подходы, модели, грамматики

# Этапы анализа текста

Лингвистический процессор –  
многоэтапный  
преобразователь  
(не показан графематический модуль: *токенизация*,  
*сегментация*)



# 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

# Этап синтаксического анализа

- Единица обработки – *предложение*
- на входе:  
результат морфологического анализа,  
на выходе:  
*синтаксическая структура предложения*
- Модель синтаксического анализа ЕЯ  
взаимосвязанно включает:
  - Способ представления синтаксической структуры предложения/фразы
  - Способ описания грамматических правил
  - Метод/алгоритм синтаксического анализа

# Основные подходы к моделированию син. анализа

- Лингвистика: центральный вопрос синтаксиса – вопрос о структуре предложения и связях (синтаксических) между словами.
- Существующие модели отличаются в основном:
  - *синтаксическими единицами*
  - *синтаксическими связями* между ними
- Общее: *синтаксическое дерево* предложения,
- Основные модели СА, синтаксическая структура:
  - *Структуры (деревья) зависимостей*
  - *Системы (деревья) составляющих*
  - + в грамматике русского языка: *теория членов предложения*

# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

*Мама мыла раму*

# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

*Мама мыла раму*

1	<i>Мама</i>	подлежащее
2	<i>мыла</i>	сказуемое
3	<i>раму</i>	прямое дополнение


# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

Как формально  
интерпретировать???

*Мама мыла раму*

1	<i>Мама</i>	<u>подлежащее</u>
2	<i>мыла</i>	<u>сказуемое</u>
3	<i>раму</i>	<u>прямое дополнение</u>





# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

Первый вариант  
формального метаязыка:

*Мама мыла раму*

		Объединено в группу вместе с:
1	<i>Мама</i>	<i>(мыла + раму)</i>
2	<i>мыла</i>	<i>раму</i>
3	<i>раму</i>	<i>мыла</i>

# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

Первый вариант  
формального метаязыка:

**Структура составляющих**

*(Мама (мыла раму))*

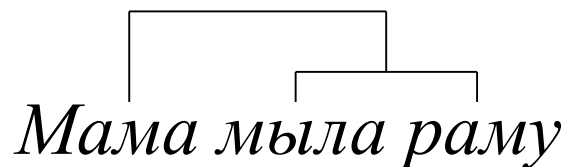
		Объединено в группу вместе с:
1	<i>Мама</i>	<i>(мыла + раму)</i>
2	<i>мыла</i>	<i>раму</i>
3	<i>раму</i>	<i>мыла</i>

# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

Первый вариант  
формального метаязыка:

**Структура составляющих**

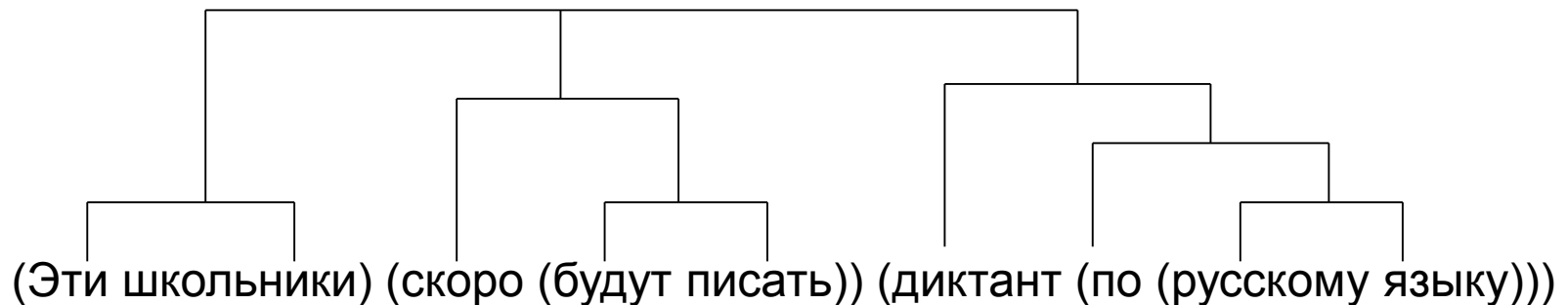


		Объединено в группу вместе с:
1	<i>Мама</i>	<i>(мыла + раму)</i>
2	<i>мыла</i>	<i>раму</i>
3	<i>раму</i>	<i>мыла</i>

# СТРУКТУРА СОСТАВЛЯЮЩИХ

## неформальное определение

- Составляющие – общее название для отдельных слов и групп в предложении, где группы – это отрезки предложения разной длины, которые объединяют более тесно связанные друг с другом единицы меньшего размера (тоже группы или отдельные слова)



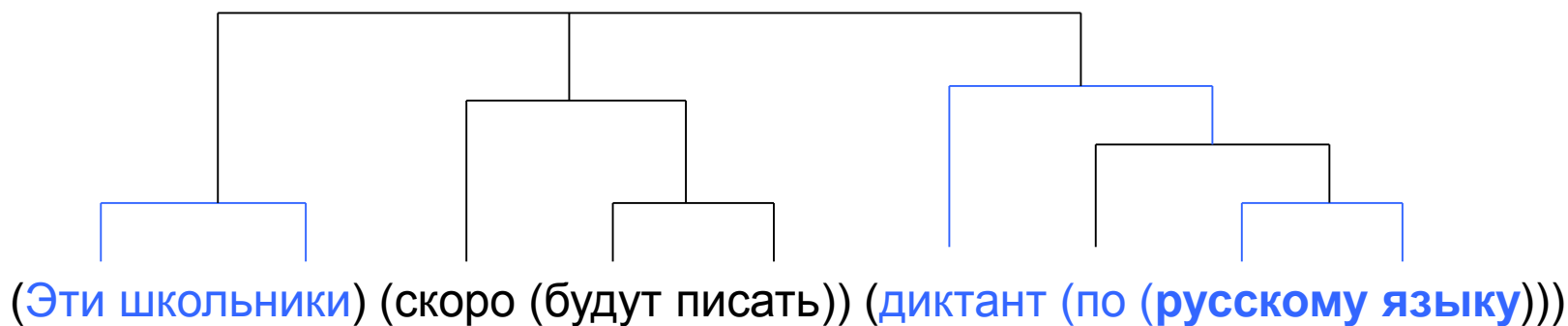
# Система составляющих: формализация

- Предложение – цепочка словоформ  $S = (w_1, w_2, \dots, w_N)$ , т.е. конечное линейное упорядоченное множество
- *Составляющая* – произвольная подпоследовательность (отрезок) цепочки
- *Система составляющих* – это такое множество  $C$  отрезков этого множества  $S$ , которое удовлетворяет следующим условиям:
  1.  $\forall w \in S : w \in C$
  2.  $S \in C$  (т.е. само предложение является элементом системы своих составляющих)
  3.  $\forall \alpha, \beta$ , таких что  $\alpha \in S, \beta \in S$   
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)

# Размеченные системы составляющих

## мотивировка

- В примере ниже: целесообразно отразить то общее, что есть между составляющими {Эти школьники}, {диктант по русскому языку}, {русскому языку} путем отнесения их к одному классу
- После этого можно определить: по каким правилам составляющие одного класса складываются из составляющих других классов? (вопрос собственно о структуре)



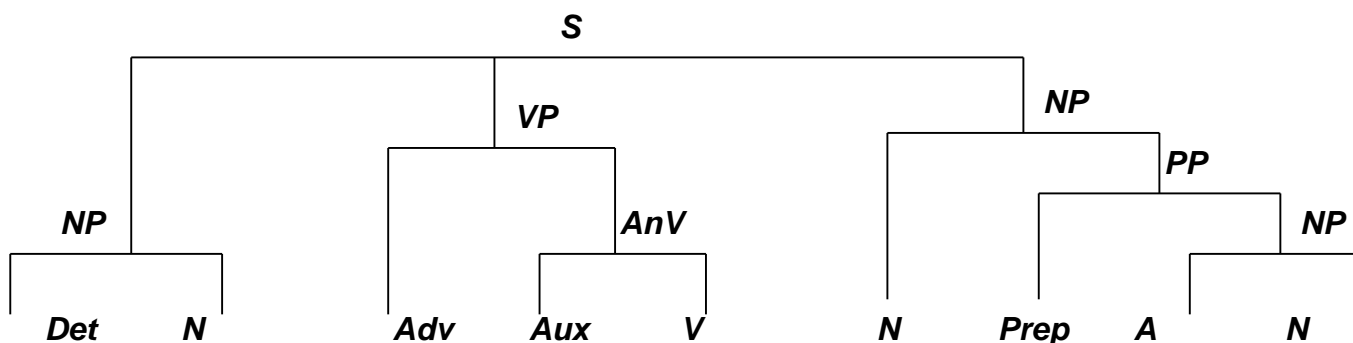
# Размеченные системы составляющих

- Размеченная система составляющих –упорядоченная тройка  $\langle C, W, \varphi \rangle$ , где
  - $C$  – система составляющих,
  - $W$  – множество меток  
(список классов, введенных в данной классификации, иначе называемых «фразовые категории»),
  - $\varphi$  – отображение  $C$  в множество всех непустых подмножеств  $W$   
(список пар «составляющая + метка/метки, приписанные данной составляющей»).

# Размеченные системы составляющих

## пример 1

W =	{S – предложение	Det – местоименное прилагательное
	NP – именная группа	N – имя существительное
	VP – глагольная группа	Adv – наречие
	AnV – аналитическая форма глагола	Aux – вспомогательный глагол
	PP – предложная группа	V – глагол
		Prep – предлог
		A – имя прилагательное}



(Эти школьники) (скоро (будут писать)) (диктант (по (русскому языку)))



# Размеченные системы составляющих

## пример 1

Эти, Det

школьники, N

Эти школьники, NP

скоро, Adv

будут, Aux

писать, V

будут писать, AnV

скоро будут писать, VP

диктант, N

по, Prep

русскому, A

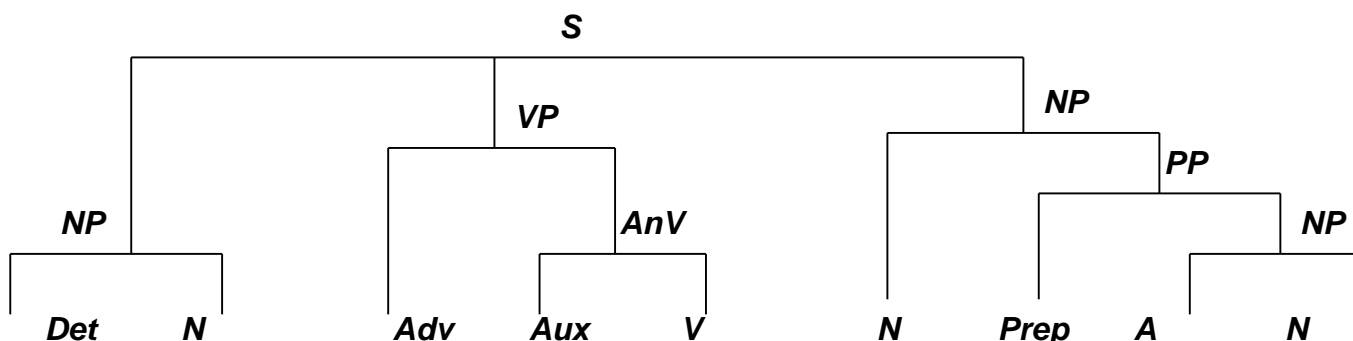
языку, N

русскому языку, NP

по русскому языку, PP

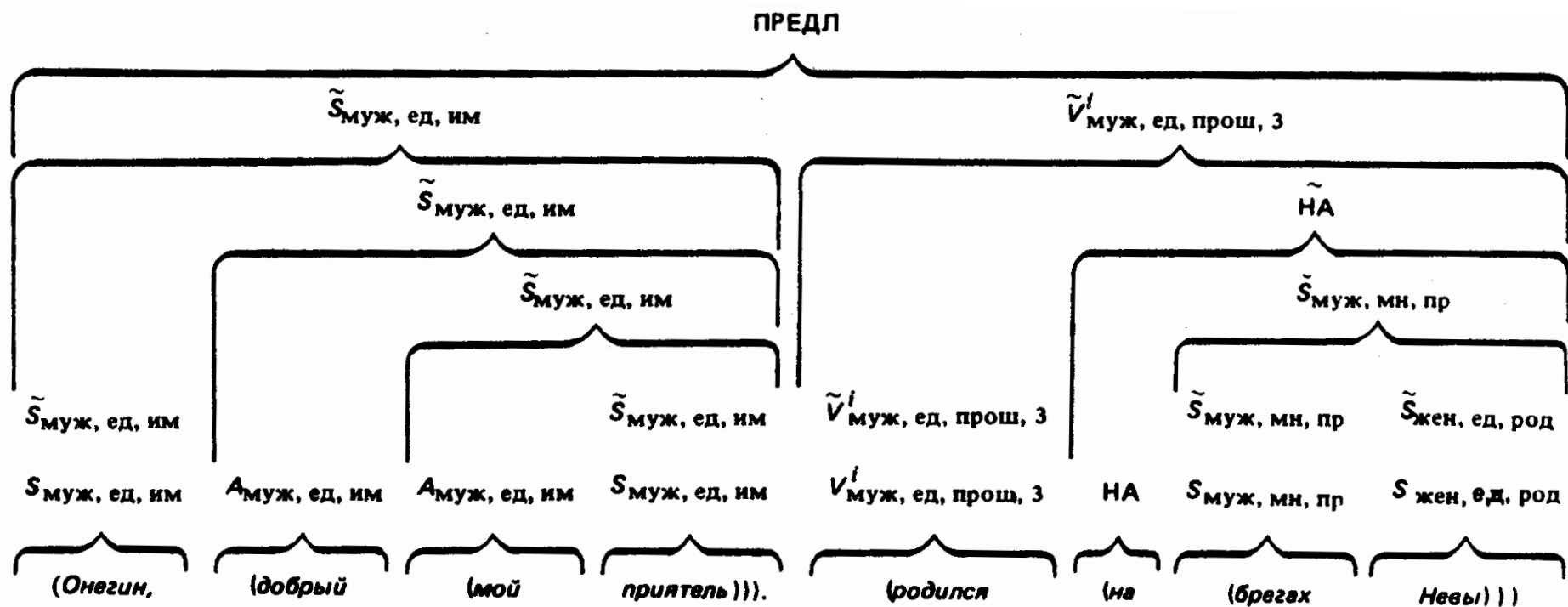
Эти школьники скоро будут писать

диктант по русскому языку, S



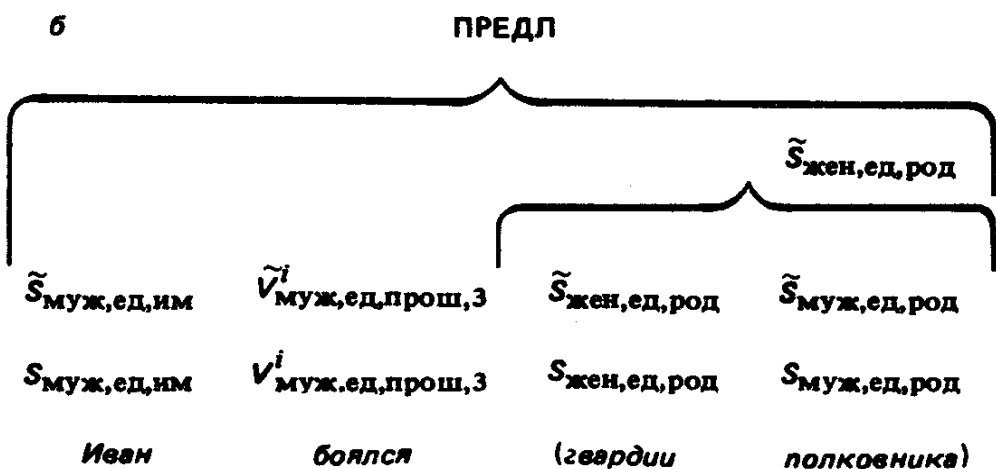
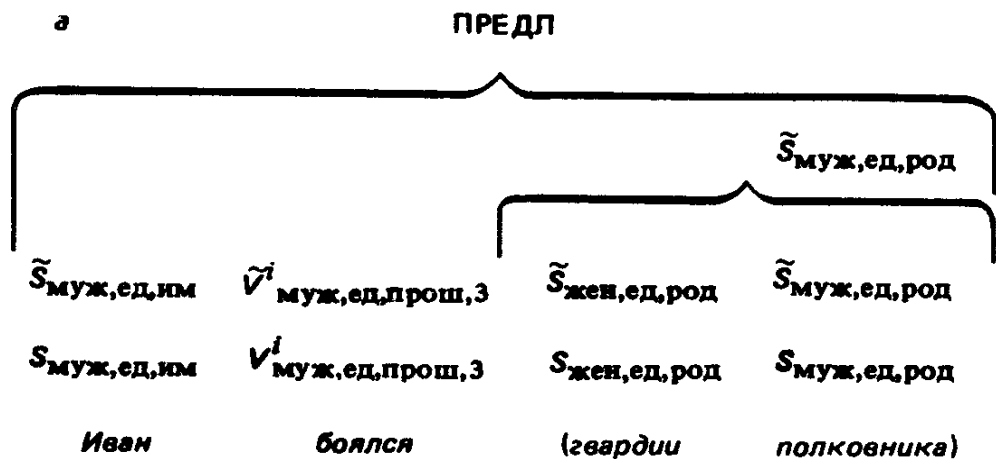
(Эти школьники) (скоро (будут писать)) (диктант (по (русскому языку)))

# Дерево составляющих: пример 2



# Размеченные системы составляющих

## пример 3



# Формализация правильной структуры

- Соответствующая грамматика должна фиксировать правильные в лингвистическом смысле фразы
- *КС-грамматика* для примера:
  - $S \rightarrow NP VP$   $S, NP, VP, A, N, Det, \dots$  –
  - $NP \rightarrow A N \mid Det N \mid N \mid N PP$  *нетерминалы*
  - $VP \rightarrow V \mid Adv V \mid V NP$   
 $AnV \mid Adv AnV \mid AnV NP$
  - $AnV \rightarrow Aux V$
  - $PP \rightarrow Prep NP$
- Нетерминалы соответствуют типам (меткам) фраз и обозначениям (меткам) частей речи слов

# Формализация синтаксической структуры

- Системы составляющих
- **Деревья зависимостей**

# Формальный подход к организации синтаксического анализа

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

Второй вариант  
формального метаязыка:

*Мама мыла раму*

		Зависит от :
1	<i>Мама</i>	<i>мыла</i>
2	<i>мыла</i>	—
3	<i>раму</i>	<i>мыла</i>

# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

Второй вариант  
формального метаязыка:

**Структура зависимостей**

  
*Мама      мыла      раму*

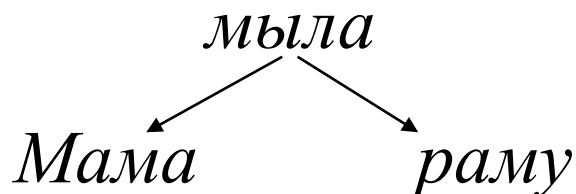
		Зависит от :
1	<i>Мама</i>	<i>мыла</i>
2	<i>мыла</i>	—
3	<i>раму</i>	<i>мыла</i>

# Формализация синтаксического представления предложения

- Мы хотим наши знания о синтаксисе формализовать. А каким метаязыком мы можем при этом пользоваться?

Второй вариант  
формального метаязыка:

**Структура зависимостей**  
**=дерево подчинения**



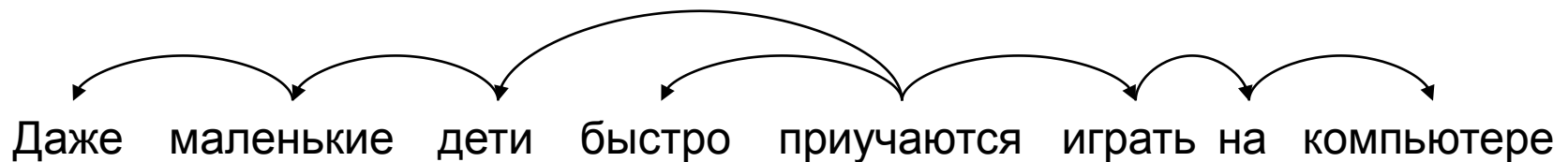
		Зависит от :
1	<i>Мама</i>	<i>мыла</i>
2	<i>мыла</i>	—
3	<i>раму</i>	<i>мыла</i>



# СТРУКТУРА ЗАВИСИМОСТЕЙ

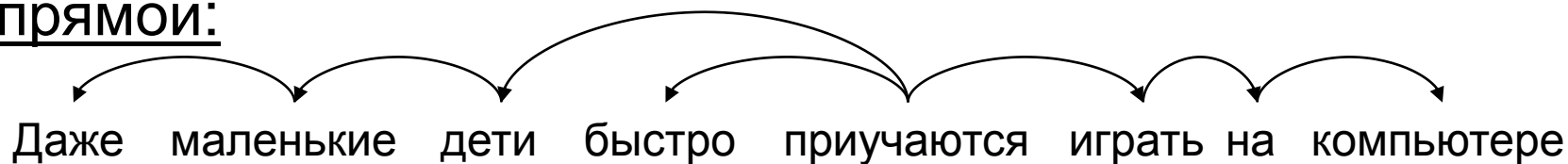
## неформальное определение

- Синтаксическая зависимость = синтаксическая связь:
  - бинарное иерархическое (формально: антисимметричное) отношение между отдельными элементами (словами в предложении);
  - антитранзитивное отношение, хотя можно говорить об опосредованном подчинении;
  - связность полной структуры предложения.



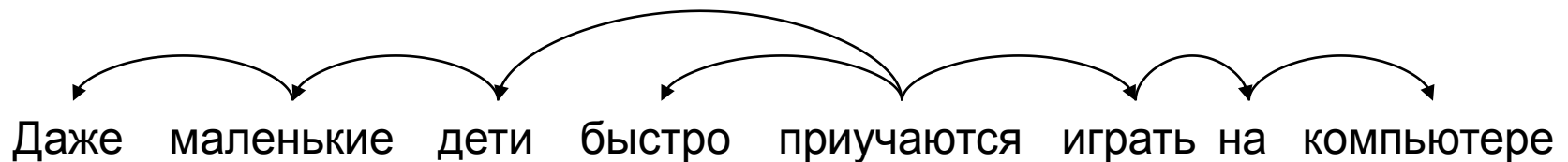
# Деревья зависимостей

- Дерево зависимостей (дерево подчинения):
  - ✓ узлы – слова предложения (корень дерева – глагол)
  - ✓ дуги – подчинительная связь (зависимость)
- Особенность: дерево предложения должно быть дополнено информацией о линейной структуре (т.е. задан порядок слов) – в отличие от систем составляющих, отражающих одновременно синтаксическую и линейную структуру предложения.
- Дерево зависимостей часто изображается в виде точек на прямой, между которыми проведены направленные дуги (зависимости); причем все дуги по одну сторону от прямой:



# Деревья зависимостей: проективность

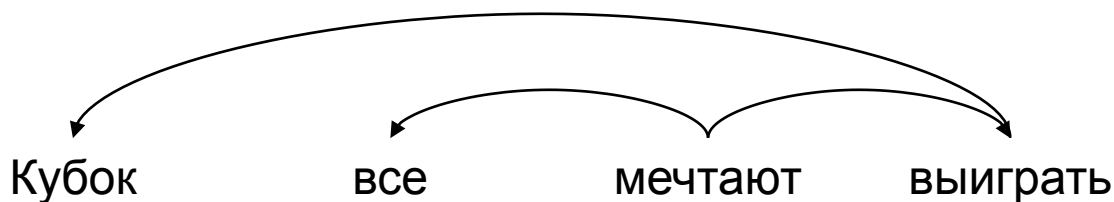
- Проективность: некая «правильность» фразы  $S$ .
- Дерево зависимостей  $\langle S, R \rangle$  для цепочки  $T$  называется *проективным*, если для  $\forall \alpha, \beta, \gamma$  - точек цепочки, таких что  $\alpha \rightarrow \beta$  и  $\gamma$  находится между  $\alpha$  и  $\beta$ , следует, что  $\gamma$  зависит от  $\alpha$ :  $\alpha \rightarrow \gamma$
- Графически *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

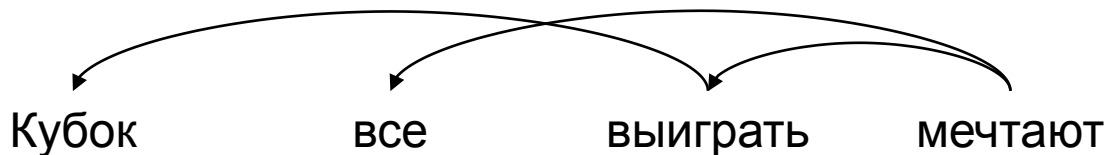
- Содержательный смысл проективности: синтаксически связанные слова близки к друг другу в цепочке слов предложения.
- Большинство правильных предложений русского языка проективны.
- Однако возможен также случай слабой проективности.
- Дерево *слабопроективно*, если ни одна из дуг не пересекает другую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

Пример слабой проективности:



# Деревья зависимостей: непроективность

- Непроективные предложения встречаются в художественной литературе, в разговорной речи в языках со свободным (или относительно свободным) порядком слов.
- Усложняют синтаксический анализ.
- Примеры:
  - *Я памятник себе воздвиг нерукотворный*
  - *Кубок все выиграть мечтают*



# Размеченные деревья зависимостей

Для анализа ЕЯ обычно используется

*размеченное дерево подчинения :*

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,

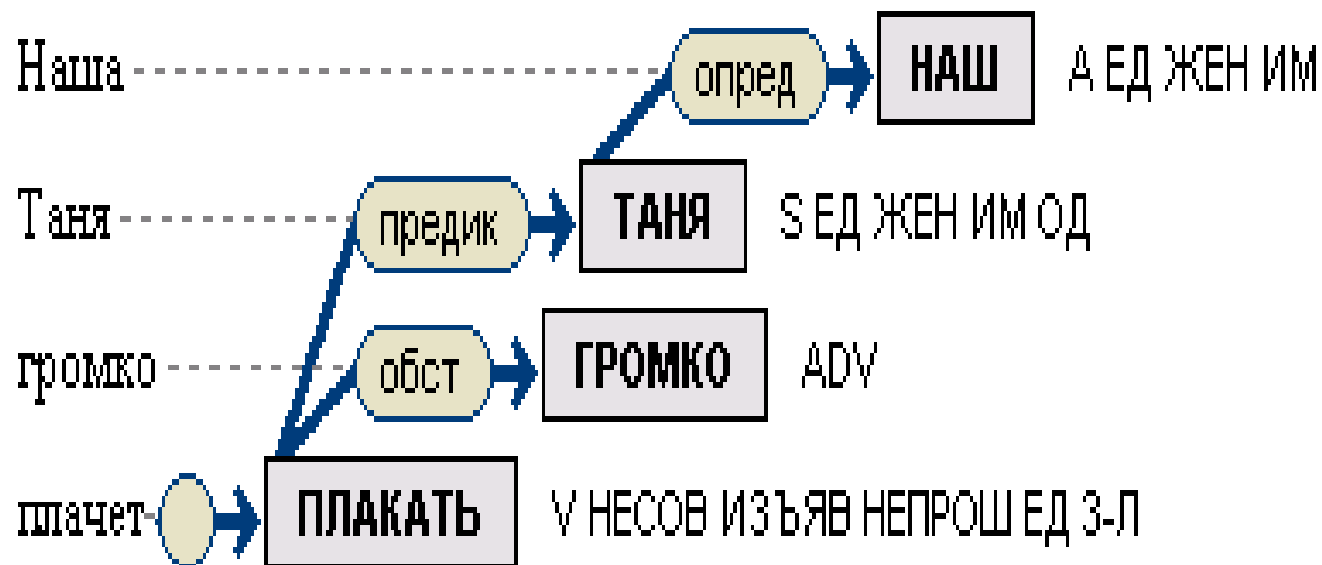
$\varphi$  – отображение множества дуг дерева во  
множество  $W$ ,

т.е. список пар «дуга дерева + метка типа  
связи».

# Типы синтаксических связей

- Набор типов зависит от модели синтаксического анализа
- Наиболее распространенные типы:
  - 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*
  - обстоятельство: *лежать* → *на* → *полу*
  - Ограничительное: *не* → *для* → *всех*

# Синтаксическое дерево





# Дерево зависимостей: главное слово и зависимое

- Критерии выбора главного слова  $H$  и зависимого слова  $D$  в конструкции  $C$ 
  - $H$  определяет синтаксическую категорию  $C$ ,  $H$  может заменить  $C$
  - $H$  определяет семантическую категорию  $C$
  - Форма  $D$  зависит от  $H$  (управление или согласование)
  - Линейная позиция  $D$  определяется по отношению к  $H$  и др.

# Очевидные случаи определения главного слова

Главное слово (Н)	Зависимое слово (D)
Глагол (сказуемое)	Существительное (подлежащее)
Глагол (сказуемое)	Существительное (дополнение)
Существительное	Прилагательное

# Неочевидные случаи определения главного слова

- Конструкция с вспомогательным глаголом (*будет читать*)
- Предложные группы (*в дом*)
- Однородные члены предложения (*белый и красный*)
- Tricky cases

# Сравнение моделей: общие свойства

- Для конкретного предложения только некоторые (размеченные) деревья составляющих и некоторые (размеченные) деревья зависимостей правильны с лингвистической точки зрения  $\Rightarrow$  для этого нужны соответствующие **грамматики**:
  - Грамматики составляющих (например, *КС-грамматики*)
  - Грамматики зависимостей
- Для отдельных предложений ЕЯ возможно более одной правильной синтаксической структуры – это случай *синтаксической омонимии*:
  - *Flying planes may be dangerous.*
  - *Я видел его молодым... Мать любит дочь.*
- Эта омонимия во многих случаях не может быть разрешена на этапе самого СА (разные смыслы)
  - принципиальная неоднозначность грамматик ЕЯ

# Сравнение моделей синтаксической структуры: отличия

- Деревья составляющих фиксируют в явном виде словосочетания, но игнорируют связи между ними
- Деревья зависимостей отображают разнотипные направленные связи, но только между словами
- Деревья составляющих больше подходят для описания синтаксиса языков со строгим (жестким) порядком слов (английский и др.)
- Деревья зависимостей – для языков с достаточно свободным порядком слов (русский, испанский и др.)

# Сравнение моделей: структуры составляющих

## Достоинства:

- Естественное представление неподчинительных отношений: *(картонка и (маленькая собачонка))*
- Возможность описать различные фразы и словокомплексы

## Недостатки:

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

# Сравнение моделей: деревья зависимостей

Достоинства:

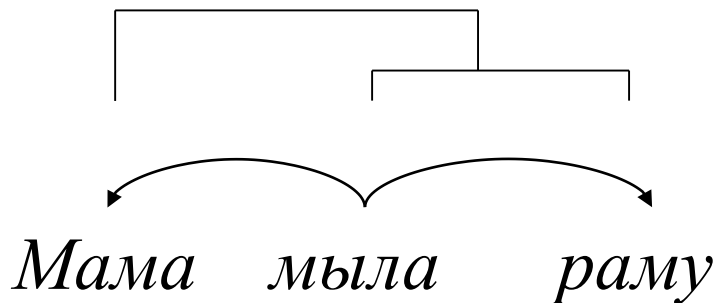
- Возможность представление непроективных структур, типизация синтаксических связей

Недостатки:

- Неоднозначности в отображении неподчинительных (сочинительных) отношений, например, однородных членов предложения: *красивый и умный*
- Не могут выразить связи разноуровневых единиц, более крупных, чем слово, например, конструкции с вспомогательным глаголом: *будет читать*
- Не позволяют выразить двойное подчинение и случаи приложений: *директор Иванов*

# ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Поиск метаязыка для описания синтаксических структур  
Как соединить сильные стороны двух рассмотренных метаязыков

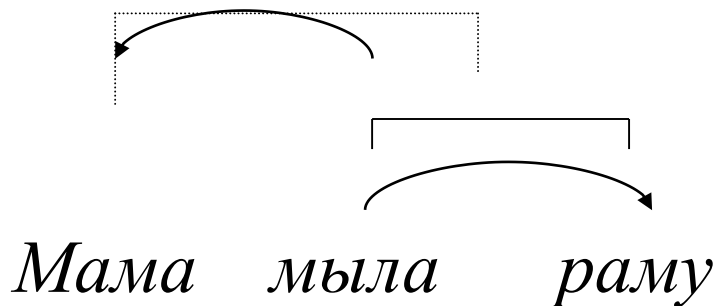


		Объединено в группу вместе с:	Зависит от :
1	<i>Мама</i>	( <i>мыла</i> + <i>раму</i> )	<i>мыла</i>
2	<i>мыла</i>	<i>раму</i>	—
3	<i>раму</i>	<i>мыла</i>	<i>мыла</i>



# ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Поиск метаязыка для описания синтаксических структур  
Как соединить сильные стороны двух рассмотренных метаязыков?



		Объединено в группу вместе с:	Зависит от :
1	Мама	(мыла + раму)	мыла
2	мыла	раму	—
3	раму	мыла	мыла

# Комбинированная модель синтаксической структуры

- Попытки преодолеть ограничения подходов
- Идея: синтаксическая структура, характеризующая не только группы слов (фразы, словосочетания), но и связность — как слов внутри групп, так и групп между собой.
- Гладкий А. (1985) — теория синтаксических групп
- понятие *синтаксической группы*: множество слов, которое вступает в отношение зависимости целиком, а не посредством одного из входящих в него слов.
- Синтаксическая группа может быть как составляющей, но может быть и разрывной.
- Сложная и многоступенчатая процедура отсеивания кандидатов в синтаксические группы.

# Комбинированная модель: пример

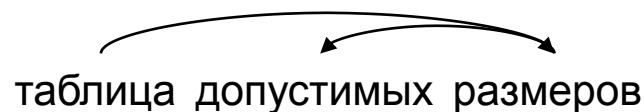
Синтаксическая группа *допустимых размеров*

(а) *таблица допустимых размеров*

(таблица, в которую сведены допустимые размеры)

(б) *таблица допустимых размеров*

(таблица, размеры которой допустимы)



# Комбинированные структуры составляющих и зависимостей

*Он любил ходить без шапки.*



*Без шапки любил ходить Иван.*



*Без шапки Иван ходить не любил*



Синтаксические группы с внутренней иерархией и без таковой, например: отсутствие внутренней иерархии в предложном сочетании. Возможность установления подчинительной связи между группами в целом; выделение в группу единиц актуального членения (здесь – темы)

Допустимость разрывных групп

# Задание в аудитории

Сколько различных синтаксических деревьев зависимостей возможны для фраз :

- *Немцов вернулся из своей командировки на север в Москву*
- *Сплочение рабочих бригад вызвало осуждение товарища министра*