

Синтаксический анализ (syntax parsing)

Павел Браславский

ВВЕДЕНИЕ

Синтаксис

- следующий уровень «на пути к смыслу»
- теоретическое значение
- приложения:
 - машинный перевод (machine translation)
 - вопросно-ответный поиск (question answering)
 - выделение информации (information extraction)
 - диалоговые системы

Особенности русского языка

- Леша подарил цветы Наташе
 - Наташе подарил цветы Леша
 - Наташа подарила цветы Леше
-
- Alice bought Bob flowers
 - Bob bought Alice flowers

Неоднозначность

Наталья приготовила курицу в кляре.

Алина приготовила курицу в духовке.

Георгий произнес речь по-немецки.

Алексей съел говядину по-строгановски.

Инна сделала движение рукой.

Жанна разрезала яблоко ножом.

Анастасия делает мир лучше.

Евгения печет блины лучше.

Задача

- по предложению построить синтаксическую структуру (или сказать, что предложение «неправильное» – не соответствует грамматике)

ФОРМАЛИЗМЫ

Два формализма

- составляющие (constituents, phase grammar, constituency parsing)
- деревья зависимостей (dependency trees)

Структура составляющих

- части речи – классы слов с одинаковыми грамматическими функциями
- группы – «расширение» частей речи
- иерархия
- традиционный формализм для описания английского синтаксиса

Примеры

[Однажды весною], [в час небывало жаркого заката], [в Москве], [на Патриарших прудах], [появились [два гражданина]].

Второй – [плечистый, рыжеватый, вихрастый молодой человек в заломленной на затылок клетчатой кепке] – был в ковбойке, жеваных белых брюках и в черных тапочках.

Теги для групп

Phrase Level

ADJP - Adjective Phrase.

ADVP - Adverb Phrase.

CONJP - Conjunction Phrase.

FRAG - Fragment.

INTJ - Interjection. Corresponds approximately to the part-of-speech tag UH.

LST - List marker. Includes surrounding punctuation.

NAC - Not a Constituent; used to show the scope of certain prenominal modifiers within an NP.

NP - Noun Phrase.

NX - Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.

PP - Prepositional Phrase.

PRN - Parenthetical.

PRT - Particle. Category for words that should be tagged RP.

QP - Quantifier Phrase (i.e. complex measure/amount phrase); used within NP.

RRC - Reduced Relative Clause.

UCP - Unlike Coordinated Phrase.

VP - Verb Phrase.

WHADJP - *Wh*-adjective Phrase. Adjectival phrase containing a *wh*-adverb, as in *how hot*.

WHAVP - *Wh*-adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a *wh*-adverb such as *how* or *why*.

WHNP - *Wh*-noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some *wh*-word, e.g. *who*, *which book*, *whose daughter*, *none of which*, or *how many leopards*.

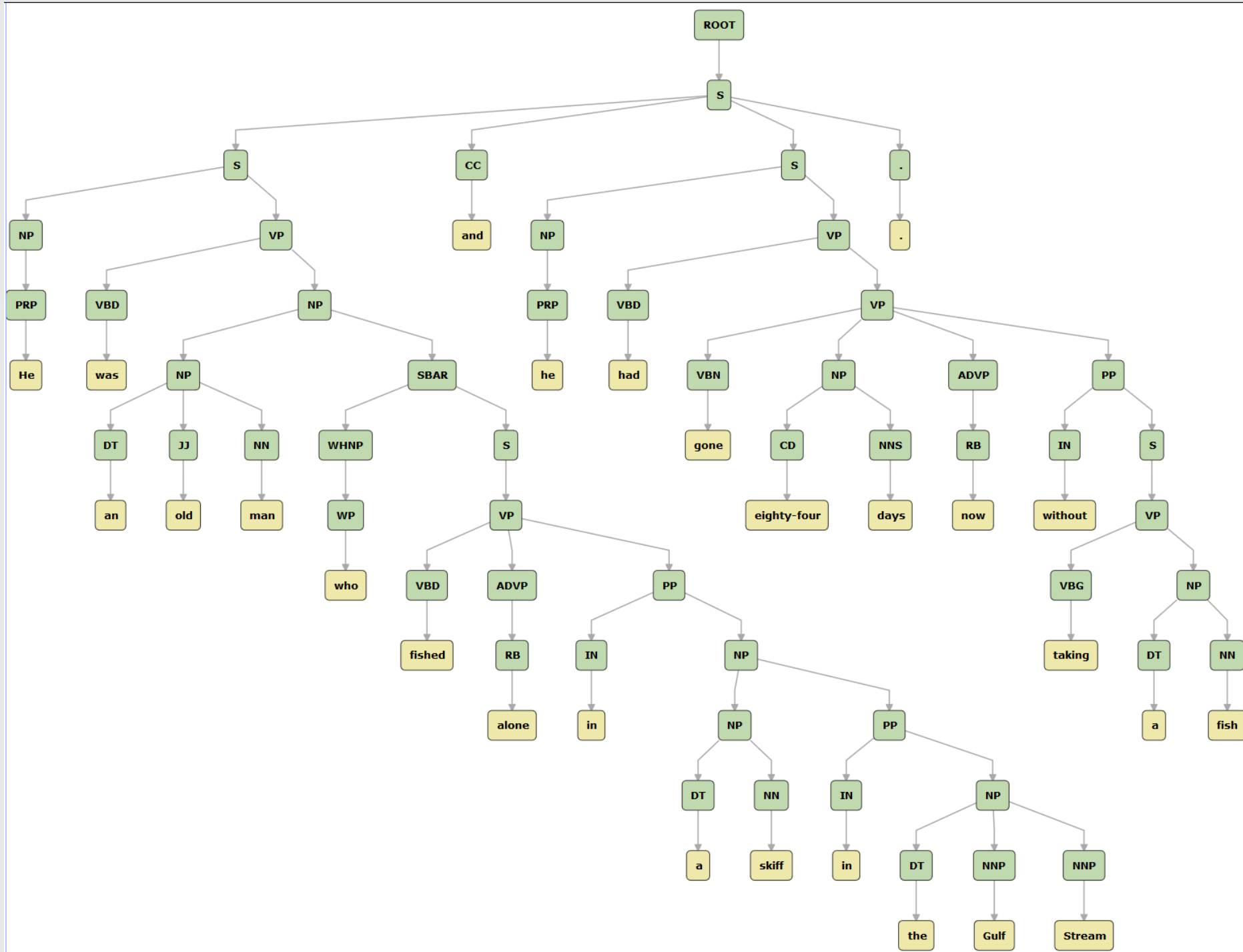
WHPP - *Wh*-prepositional Phrase. Prepositional phrase containing a *wh*-noun phrase (such as *of which* or *by whose authority*) that either introduces a PP gap or is contained by a WHNP.

X - Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing *the...the*-constructions.

Тип	Сокращенное название	Пример
Количественная группа	КОЛИЧ	двадцать восемь
Последовательность чисел вперемешку со знаками препинания	КОЛИЧ	12,2
Существительное из заданного перечня + числовой идентификатор	СУЩ-ЧИСЛ	статья 123
Правила для построения ФИО (используются морфологические пометы о том, что данное слово может быть именем)	ФИО	Петров Петр Владимирович
Слова степени (типа "очень") с группой прилагательного или причастия	НАР_ПРИЛ	очень красивый
Однородные прилагательные	ОДНОР_ПРИЛ	первой и единственной
Однородные наречия	ОДНОР_НАР	долго иль коротко
Однородные инфинитивы	ОДНОР_ИНФ	стоять или лежать
Однородные прилагательные сравнительной степени	ОДНОР_ПРИЛ	красивее и моложе
Группы даты	ДАТА	август 1968 года, 12 июня 99 г. и т.д.
Группа временных отрезков	СЛОЖ_ПГ	С первого августа по двадцатое сентября
Аналитическая форма сравнительной степени прил. или наречия	СРАВН-СТЕПЕНЬ	гораздо сильнее
Наречие + глагол	НАРЕЧ-ГЛАГОЛ	злостно нарушать
Одно или несколько прилагательных, согласованных по роду, числу и падежу со стоящим сразу после них существительным.	ПРИЛ-СУЩ	длинная унылая дорога
Наречное числительное + ИГ (рд мн)	НАР-ЧИСЛ-СУЩ	много очень простых ребят
Числительное + ИГ	ЧИСЛ-СУЩ	сорок восемь попугаев
Генитивная пара	ГЕНИТ_ИГ	рука Москвы
Предложная группа	ПГ	на холме
Однородные ИГ	ОДНОР_ИГ	мама и папа
Отрицание + глагольная форма	ОТР_ФОРМА	не любить
Глагольная форма+контактное прямое дополнение	ПРЯМ_ДОП	рубить дрова
Группа электронного адреса	ЭЛ_АДРЕС	www.dialing.ru
Глагольная форма+контактный инфинитив	ГЛАГ_ИНФ	пойти выпить
Подлежащее	ПОДЛ	я пошел
Сказуемое	вершина клаузы	я пошел

http://aot.ru/demo/rus_syn_consts.html

<http://aot.ru/docs/synan.html>





ГРАММАТИКА ЗАВИСИМОСТЕЙ

Грамматика зависимостей

- главное/зависимое слово (хозяин/слуга)
- центральная роль личного глагола
- проективность – желательное свойство

Очень они хорошие были люди.

- в последние годы – основной формализм, в т.ч. в проекте Universal dependencies
- может быть получена из структуры составляющих

Пример



Пример

Stanford CoreNLP

— Text to annotate —

But among people who have spent years in the tech industry, Intuit is just as memorable for being the company that got away from Microsoft in the mid-1990s — a rare moment when Bill Gates's giant, at the peak of its influence, did not get its way.

— Annotations —

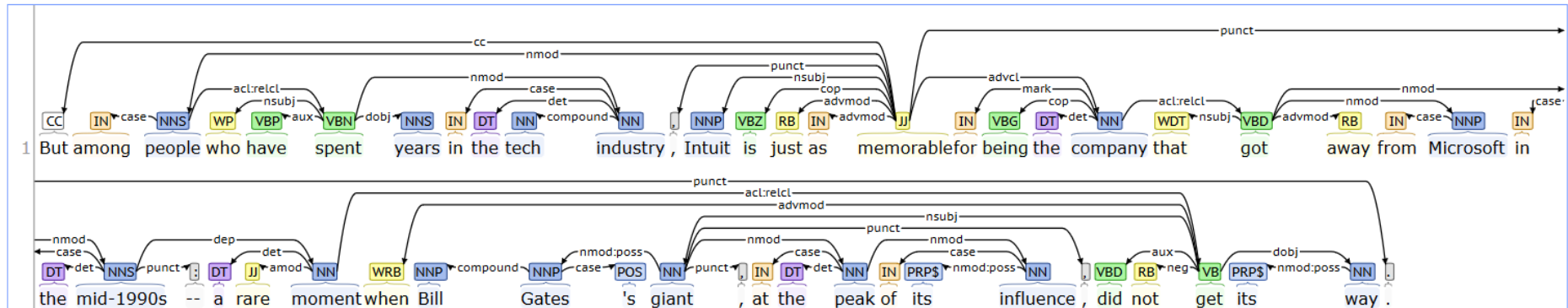
dependency parse x

— Language —

English

Submit

Basic Dependencies:



Отношения [Мельчук]

- a. субъектное: *мальчик* ← *бежит*
- b. прямое дополнение: *вижу* → *мальчика*
- c. косвенное дополнение: *сказал* → *мне*
- d. аппозиция: *диван* ← *кровать*
- e. ограничение/отрицание: *только* ← *посмотреть*
- f. определение: *первая* ← *страница*
- g. посессивное: *книга* → *девочки*
- h. количественное: *три* → *апельсина*
- i. обстоятельство: *бойко* ← *говорить*
- j. предложное: *без* → *головы*
- k. присвязочное: *был* → *болен*
- l. сочинительное: *война и* → *мир*
- m. инфинитивно-дополнительное: *хочу* → *спать*

UD: типы зависимостей

Core dependents of clausal predicates

Nominal dep	Predicate dep	
<u>nsubj</u>	<u>csbj</u>	
<u>nsubjpass</u>	<u>csbjpass</u>	
<u>dobj</u>	<u>ccomp</u>	<u>xcomp</u>
<u>iobj</u>		

Noun dependents

Nominal dep	Predicate dep	Modifier word
<u>nummod</u>	<u>acl</u>	<u>amod</u>
<u>appos</u>		<u>det</u>
<u>nmod</u>		<u>neg</u>

Case-marking, prepositions, possessive

[case](#)

Non-core dependents of clausal predicates

Nominal dep	Predicate dep	Modifier word
<u>nmod</u>	<u>advcl</u>	<u>advmod</u>
		<u>neg</u>

Compounding and unanalyzed

<u>compound</u>	<u>mwe</u>	<u>goeswith</u>
<u>name</u>	<u>foreign</u>	

Loose joining relations

<u>list</u>	<u>parataxis</u>	<u>remnant</u>
<u>dislocated</u>		<u>reparandum</u>

Special clausal dependents

Nominal dep	Auxiliary	Other
<u>vocative</u>	<u>aux</u>	<u>mark</u>
<u>discourse</u>	<u>auxpass</u>	<u>punct</u>
<u>expl</u>	<u>cop</u>	

Coordination

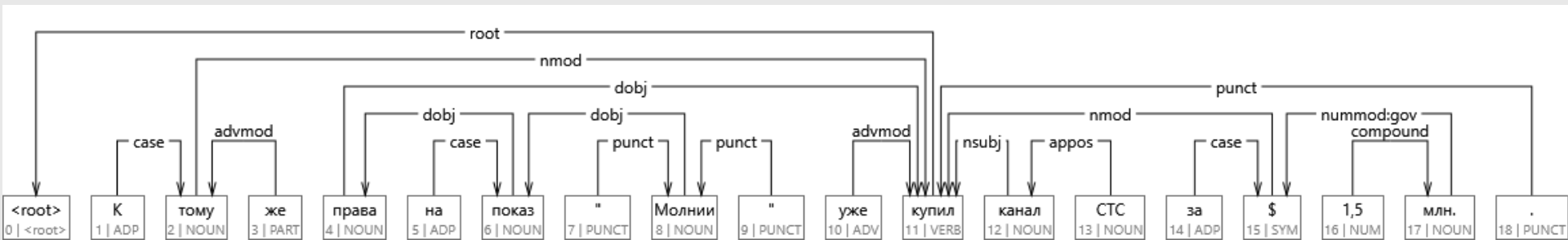
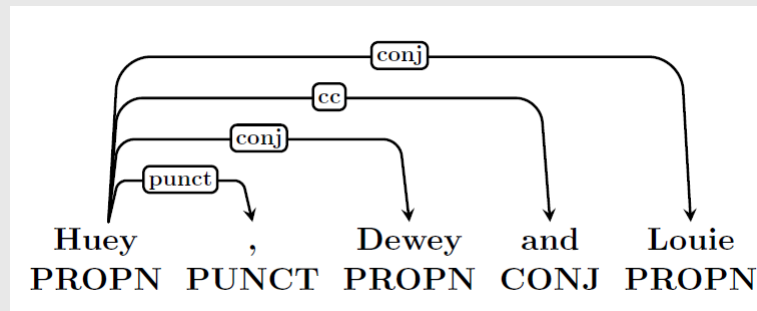
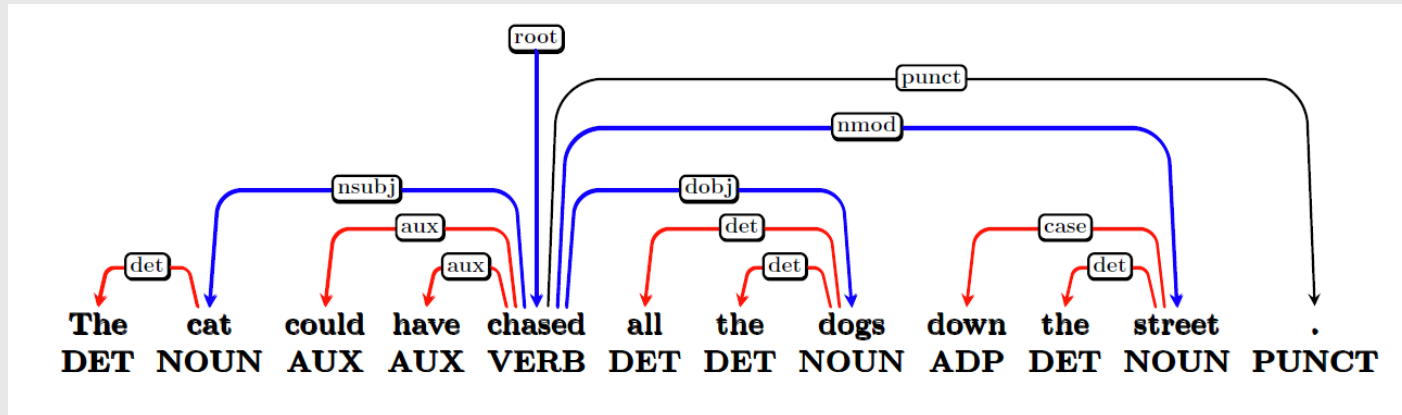
<u>conj</u>	<u>cc</u>	<u>punct</u>
-----------------------------	---------------------------	------------------------------

Other

Sentence head	Unspecified dependency
<u>root</u>	<u>dep</u>

<http://universaldependencies.org/u/dep/index.html>

UD: пример



ДАННЫЕ

Penn Treebank

- ~ 40K предложений / 2,400 тестовых
- Wall Street Journal + устная речь
- составляющие

<http://www.cis.upenn.edu/~treebank/home.html>

The Penn Treebank

```
( (S
  (NP-SBJ (DT The) (NN move))
  (VP (VBD followed)
    (NP
      (NP (DT a) (NN round))
      (PP (IN of)
        (NP
          (NP (JJ similar) (NNS increases))
          (PP (IN by)
            (NP (JJ other) (NNS lenders))))
          (PP (IN against)
            (NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
    (, ,)
    (S-ADV
      (NP-SBJ (-NONE- *))
      (VP (VBG reflecting)
        (NP
          (NP (DT a) (VBG continuing) (NN decline))
          (PP-LOC (IN in)
            (NP (DT that) (NN market))))))
      (. .)))
```


SynTagRus

- 64K предложений (2015)
- жанры: художественный, научно-популярный, публицистический, биографический и новостной
- 67 типов отношений
- ЗАВИСИМОСТИ

<http://ruscorpora.ru/instruction-syntax.html>

https://github.com/UniversalDependencies/UD_Russian-SynTagRus



Результаты поиска в синтаксическом корпусе

[перейти на страницу](#)

Объем всего корпуса: 612 документов, 59 240 предложений, 860 720 слов.

"корабль"

Найдено 12 документов, 44 предложения, 44 вхождения.

Поискать в других корпусах: [основном](#), [акцентологическом](#), [газетном](#), [диалектном](#), [мультимедийном](#), [обучающем](#), [параллельном](#), [поэтическом](#), [устном](#).

Страницы: 1 [2](#) [следующая страница](#)

1. Советский шаттл [Все примеры \(4\)](#)

Пятнадцать лет назад, 15 ноября 1988 года, совершил свой полет, закончившийся не повторенной до сих пор автоматической посадкой на посадочную полосу космический **корабль** "Буран". [\[Советский шаттл\]](#) [\[Показать структуру\]](#)

Хотя для первого космонавта наряду с "гагаринским" "Востоком" ОКБ- 256 Павла Цыбина проектировало крылатый космический **корабль** классической (Космический Аппарат). [\[Советский шаттл\]](#) [\[Показать структуру\]](#)

Предполагалось, что после запуска и работы на орбите **корабль** входит в плотные слои атмосферы и совершает управляемый спуск и парашютную посадку с помощью двигателей мягкой посадки. [\[Советский шаттл\]](#) [\[Показать структуру\]](#)

Оставалось только сделать **корабль** и носитель. [\[Советский шаттл\]](#) [\[Показать структуру\]](#)

2. Космонавт Гречко [Все примеры \(15\)](#)

- Неужели веришь, что найдешь **корабль** инопланетян - спрашиваю у неугомонного друга. [\[Космонавт Гречко\]](#) [\[Показать структуру\]](#)

А когда работали над полетом человека в космос, как раз выбирал угол, под которым надо входить в атмосферу, чтобы посадить **корабль**. [\[Космонавт Гречко\]](#)

Один **корабль** ушел в космос, другой погиб во время спуска, третий потерпел аварию. [\[Космонавт Гречко\]](#) [\[Показать структуру\]](#)

Да и с запуском Юрия Гагарина не все соглашались, некоторые считали, что нужно еще проверить **корабль** в реальном полете. [\[Космонавт Гречко\]](#) [\[Показать структуру\]](#)

В свое время в одной из повестей я написал, что ты искал там **корабль** инопланетян и что Сергей Павлович Королев очень интересовался этой экспедицией.

- Еще юношей я прочитал статью писателя Казанцева о том, что Тунгусский метеорит - это космический **корабль**, который потерпел катастрофу. [\[Космонавт Гречко\]](#)

Мы использовали отчет Золотова, страниц на пятьсот, и из него следовало, что это был космический **корабль**, который взорвался на высоте пять километров. [\[Космонавт Гречко\]](#) [\[Показать структуру\]](#)

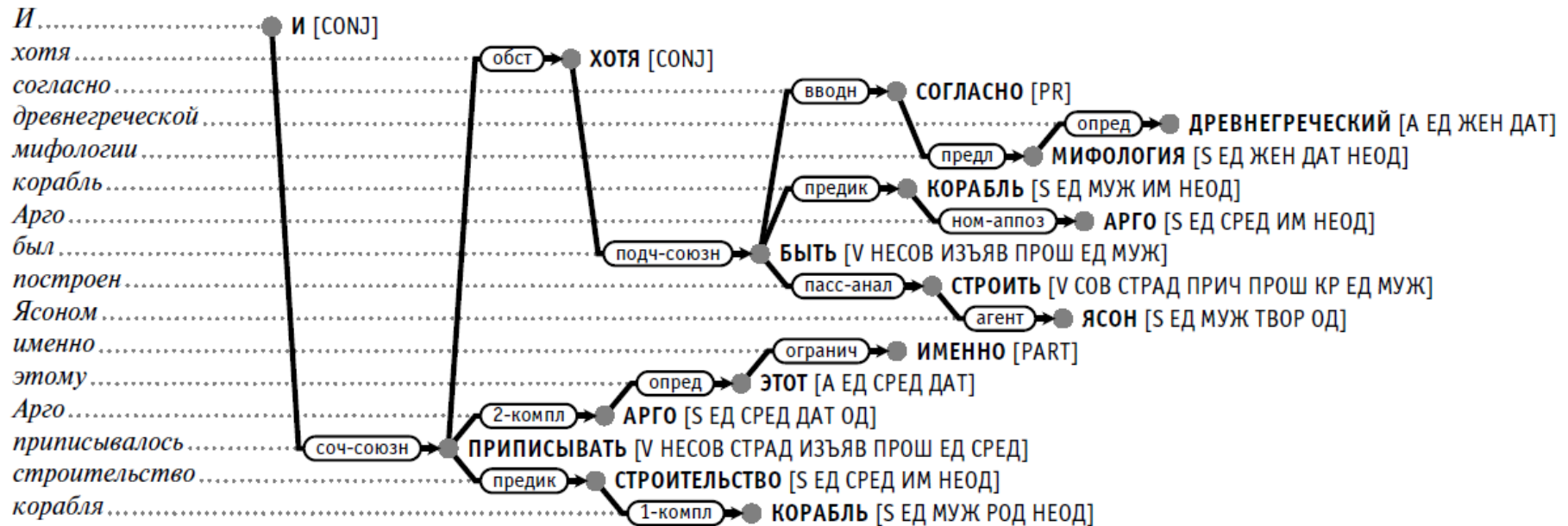
У Константина Петровича было какое-то удивительное чутье, он без данных, без расчетов мог сконструировать **корабль**. [\[Космонавт Гречко\]](#) [\[Показать структуру\]](#)

Он считает, что именно он сделал **корабль**. [\[Космонавт Гречко\]](#) [\[Показать структуру\]](#)

Но я считаю, что **корабль** сделал все-таки Королев. [\[Космонавт Гречко\]](#) [\[Показать структуру\]](#)

3. На Марс, как и себе домой [Все примеры \(8\)](#)

SynTagRus – пример



Many syntactic treebanks have been developed for a wide variety of languages:

Language ↕	Treebank ↕	Syntactic Formalism ↕	Distribution / License ↕
Arabic	Penn Arabic Treebank	Phrase structure	Linguistic Data Consortium
Arabic	Prague Arabic Dependency Treebank (PADT)	Dependency	Linguistic Data Consortium
Arabic	Columbia Arabic Treebank (CATiB) 	Dependency	Linguistic Data Consortium
Arabic (classical)	Quranic Arabic Dependency Treebank (QADT)  (Quranic Arabic Corpus)	Dependency	Open source (GNU general public license)
Bulgarian	BulTreeBank	HPSG	Freely available for research
Catalan	Cat3LB	Phrase structure	Freely available for research
Chinese	Penn Chinese Treebank	Phrase structure	Linguistic Data Consortium
Chinese	Sinica Treebank	Case grammar	Not freely available
Chinese	Chinese Dependency Treebank	Dependency	Linguistic Data Consortium
Croatian	Croatian Dependency Treebank	Dependency	Open source (Creative Commons license)
Czech	Prague Dependency Treebank	Dependency	Linguistic Data Consortium
Danish	Danish Dependency Treebank	Dependency	Open source (GNU general public license)
Danish	Arboretum: A syntactic tree corpus of Danish	Phrase structure	License fee
Dutch	Spoken Dutch Corpus (CGN)	Phrase structure	License fee
Dutch	Alpino Treebank	Dependency	Open source (GNU general public license)
Dutch	LASSY Small and Large	Dependency	License fee
English	Penn Treebank	Phrase structure	Linguistic Data Consortium
English	CCGbank	Combinatory categorial grammar	Linguistic Data Consortium
English	Prague English Dependency Treebank	Dependency	Linguistic Data Consortium
English	Universal Dependencies	Dependency	Open source (Creative Commons license or GNU general public license)
English	BLLIP WSJ corpus	Phrase structure	Linguistic Data Consortium
English	British Component of the International Corpus of English (ICE-GB)	Phrase structure	License fee 
English	Diachronic Corpus of Present-Day Spoken English (DCPSE)	Phrase structure	License fee 
English	Lancaster Parsed Corpus	Phrase structure	?
English	Susanne Corpus	Phrase structure	Freely available for research
English	Christine Corpus	Phrase structure	Freely available for research
English	Lucy Corpus	Phrase structure	Freely available for research
English	Tübingen Treebank of English / Spontaneous Speech (TüBa-E/S)	HPSG	Freely available for research
English	LinGO Redwoods	HPSG	?
English	Multi-Treebank	Phrase structure	Available online for comparison purposes
English	The PARC 700 Dependency Bank 	Dependency	?

ПАРСЕРЫ

Контекстно-свободная грамматика

- context-free grammars (CFG)
- $G = (\Sigma, N, S, R)$
 - Σ : терминалы
 - N : нетерминалы
 - S : начальный символ ($S \in N$)
 - R : правила (продукции) в виде $X \rightarrow \gamma$
 - $X \in N; \gamma \in (N \cup \Sigma)^*$

Пример грамматики

$S \rightarrow NP VP$

$VP \rightarrow V NP$

$VP \rightarrow V NP PP$

$NP \rightarrow NP NP$

$NP \rightarrow NP PP$

$NP \rightarrow N$

$NP \rightarrow e$

$PP \rightarrow P NP$

people fish tanks
people fish with rods

$N \rightarrow \text{people}$

$N \rightarrow \text{fish}$

$N \rightarrow \text{tanks}$

$N \rightarrow \text{rods}$

$V \rightarrow \text{people}$

$V \rightarrow \text{fish}$

$V \rightarrow \text{tanks}$

$P \rightarrow \text{with}$

[Chris Manning]

Нормальная форма Хомского

- Все правила в виде
 - $X \rightarrow YZ$
 - $X \rightarrow w$
 - $X, Y, Z \in N; w \in \Sigma$

Пример

$S \rightarrow NP VP$

$VP \rightarrow V NP$

$S \rightarrow V NP$

$VP \rightarrow V @VP_V$

$@VP_V \rightarrow NP PP$

$S \rightarrow V @S_V$

$@S_V \rightarrow NP PP$

$VP \rightarrow V PP$

$S \rightarrow V PP$

$NP \rightarrow NP NP$

$NP \rightarrow NP PP$

$NP \rightarrow P NP$

$PP \rightarrow P NP$

$NP \rightarrow \textit{people}$

$NP \rightarrow \textit{fish}$

$NP \rightarrow \textit{tanks}$

$NP \rightarrow \textit{rods}$

$V \rightarrow \textit{people}$

$S \rightarrow \textit{people}$

$VP \rightarrow \textit{people}$

$V \rightarrow \textit{fish}$

$S \rightarrow \textit{fish}$

$VP \rightarrow \textit{fish}$

$V \rightarrow \textit{tanks}$

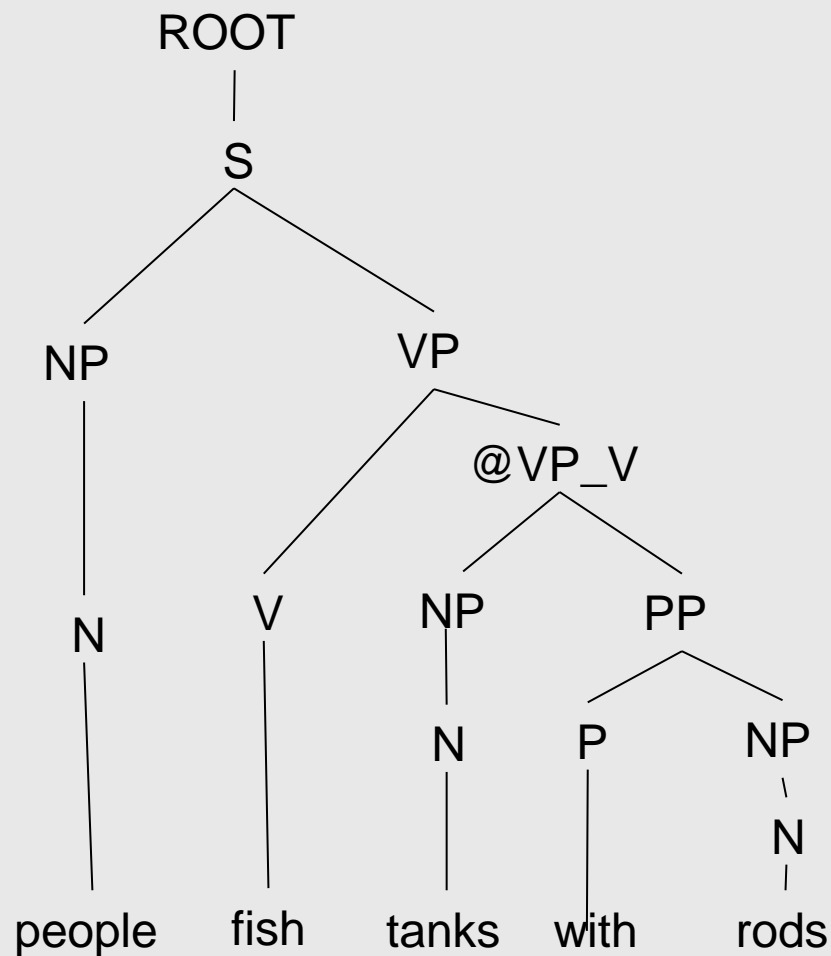
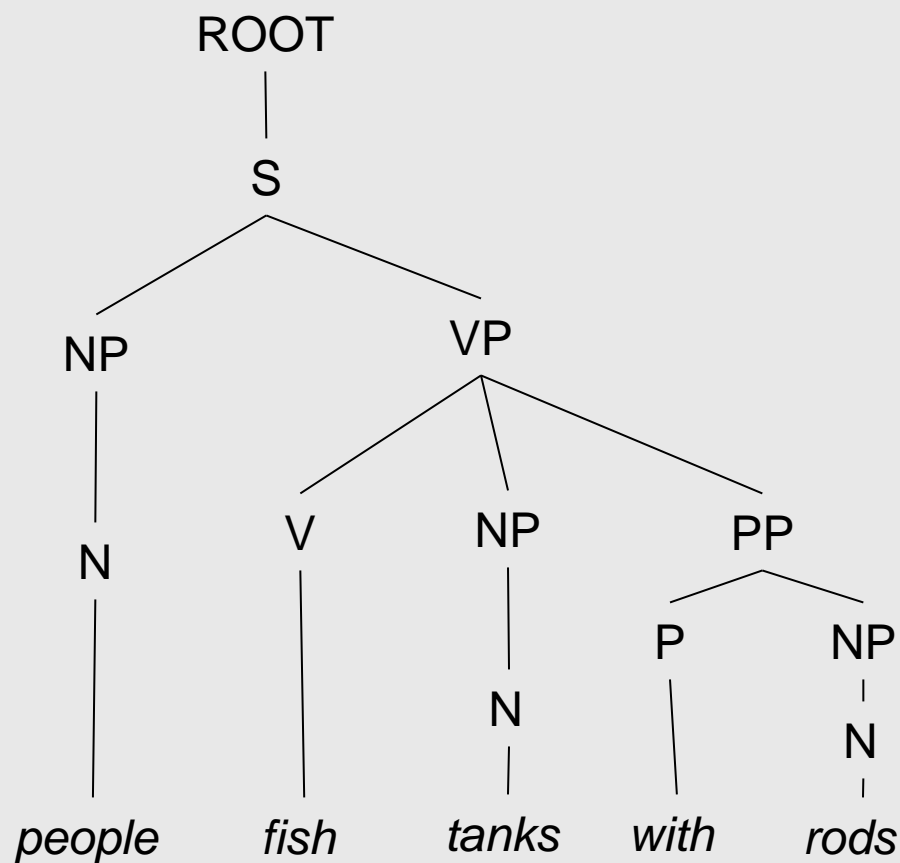
$S \rightarrow \textit{tanks}$

$VP \rightarrow \textit{tanks}$

$P \rightarrow \textit{with}$

$PP \rightarrow \textit{with}$

До и после бинаризации



Классические парсеры

- Алгоритм CYK (Cocke–Younger–Kasami)
- Демо: <http://lxmls.it.pt/2015/cky.html>
- Экспоненциальный рост количества разборов

the	DT	NP			S			S
child	N							
ate	V			VP				VP
the	DT	NP						NP
cake	N							
with	PRP							PP
the	DT	NP						
								N

[Radev]

СТАТИСТИЧЕСКИЕ ПАРСЕРЫ

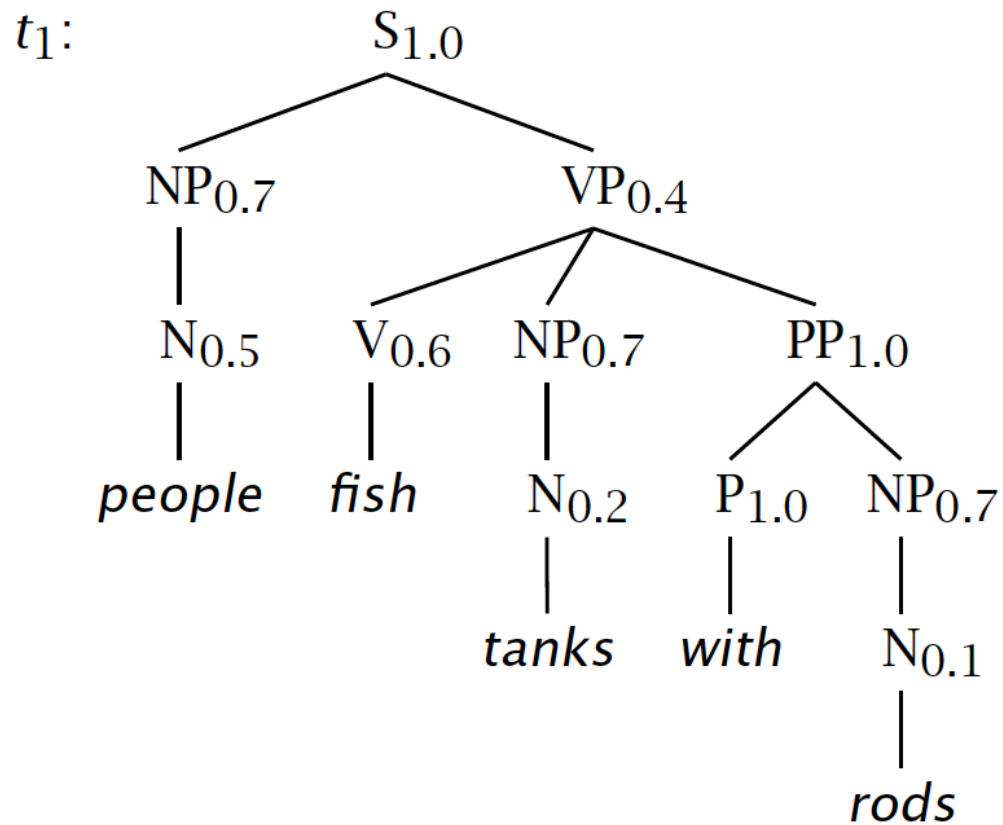
Вероятностные КСГ

- Probabilistic context-free grammars (PCFGs)
- $G = (\Sigma, N, S, R, P)$
 - Σ : терминалы
 - N : нетерминалы
 - S : начальный символ ($S \in N$)
 - R : правила (продукции) в виде $X \rightarrow \gamma$
 - P : вероятности правил
 - $P: R \rightarrow [0,1] \quad \forall X \in N, \sum_{X \rightarrow \gamma \in R} P(X \rightarrow \gamma) = 1$

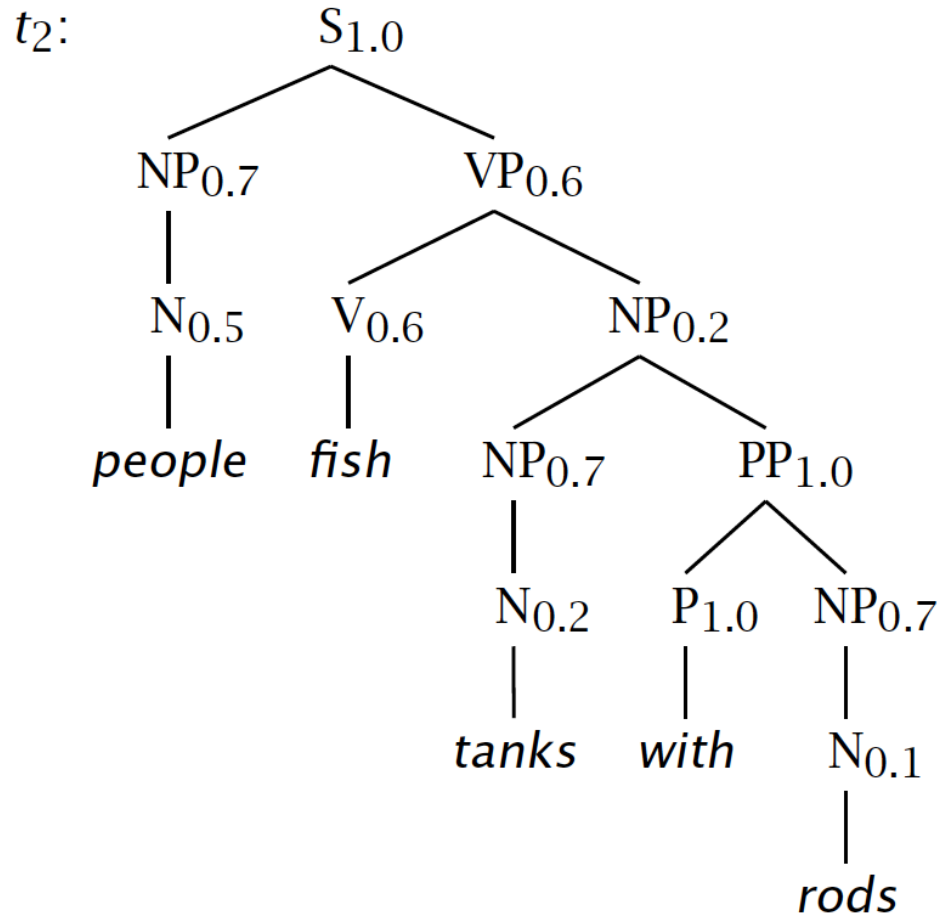
Пример

$S \rightarrow NP VP$	1.0	$N \rightarrow people$	0.5
$VP \rightarrow V NP$	0.6	$N \rightarrow fish$	0.2
$VP \rightarrow V NP PP$	0.4	$N \rightarrow tanks$	0.2
$NP \rightarrow NP NP$	0.1	$N \rightarrow rods$	0.1
$NP \rightarrow NP PP$	0.2	$V \rightarrow people$	0.1
$NP \rightarrow N$	0.7	$V \rightarrow fish$	0.6
$PP \rightarrow P NP$	1.0	$V \rightarrow tanks$	0.3
		$P \rightarrow with$	1.0

Пример – 1



Пример – 2



Пример: вероятности разбора

- $s = \textit{people fish tanks with rods}$

- $$\begin{aligned} P(t_1) &= 1.0 \times 0.7 \times 0.4 \times 0.5 \times 0.6 \times 0.7 \\ &\quad \times 1.0 \times 0.2 \times 1.0 \times 0.7 \times 0.1 \\ &= 0.0008232 \end{aligned}$$

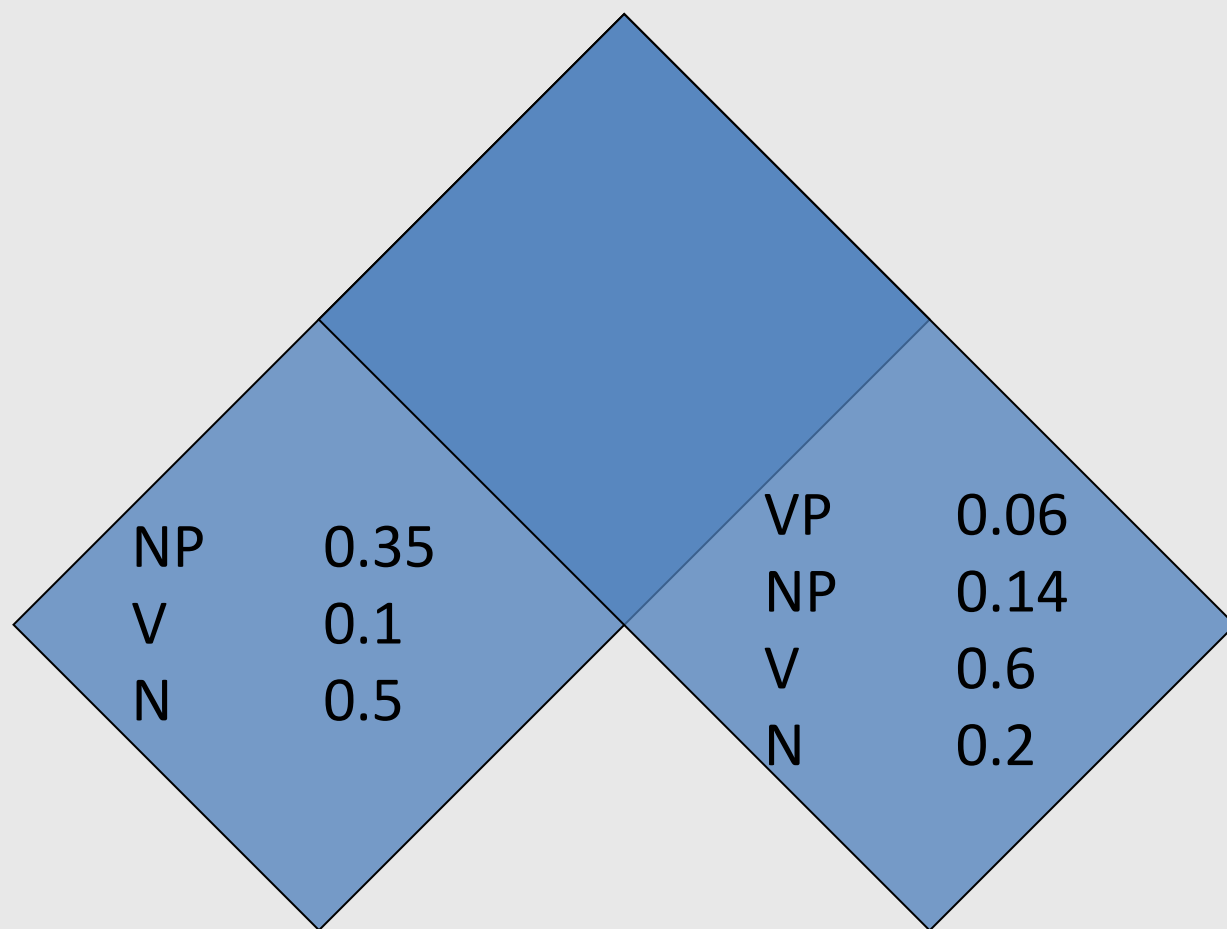
Verb attach

- $$\begin{aligned} P(t_2) &= 1.0 \times 0.7 \times 0.6 \times 0.5 \times 0.6 \times 0.2 \\ &\quad \times 0.7 \times 1.0 \times 0.2 \times 1.0 \times 0.7 \times 0.1 \\ &= 0.00024696 \end{aligned}$$

Noun attach

- $$\begin{aligned} P(s) &= P(t_1) + P(t_2) \\ &= 0.0008232 + 0.00024696 \\ &= 0.00107016 \end{aligned}$$

Viterbi (Max) Scores



people

fish

$S \rightarrow NP VP$	0.9
$S \rightarrow VP$	0.1
$VP \rightarrow V NP$	0.5
$VP \rightarrow V$	0.1
$VP \rightarrow V @VP_V$	0.3
$VP \rightarrow V PP$	0.1
$@VP_V \rightarrow NP PP$	1.0
$NP \rightarrow NP NP$	0.1
$NP \rightarrow NP PP$	0.2
$NP \rightarrow N$	0.7
$PP \rightarrow P NP$	1.0

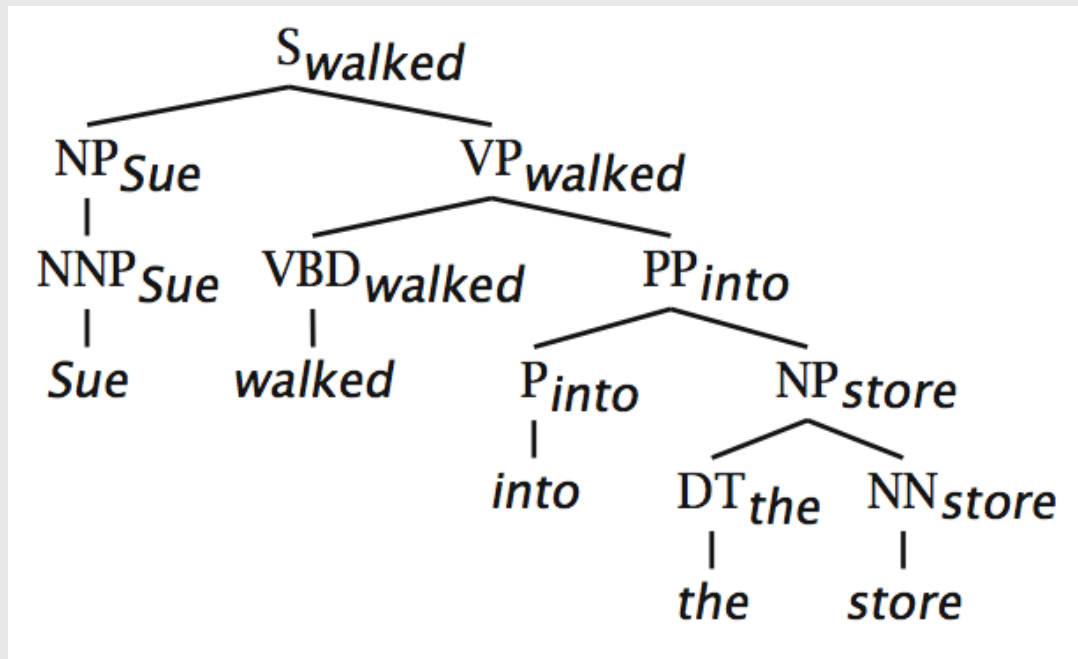
ПОВЫШЕНИЕ КАЧЕСТВА

Как можно улучшить качество?

- Лексикализация: приписываем главное слово фразы
- Учет контекста разбора
- Переопределение тегов

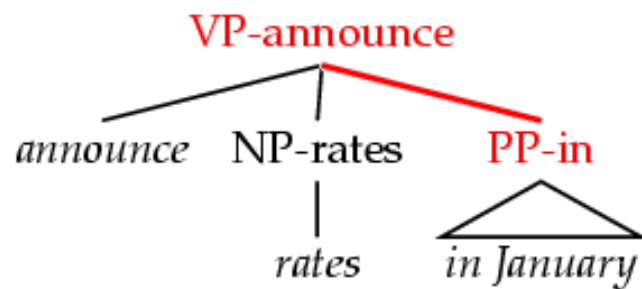
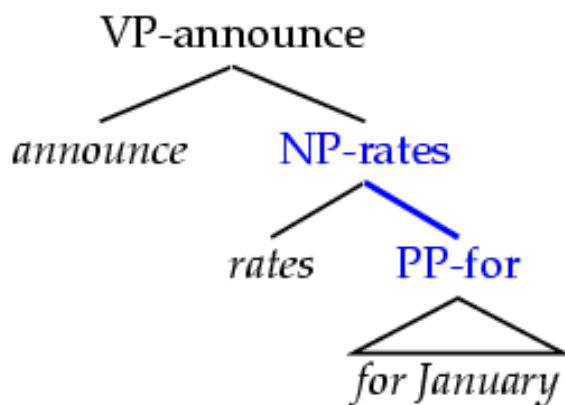
Лексикализация

[Magerman 1995, Collins 1997; Charniak 1997]



Лексикализация – 2

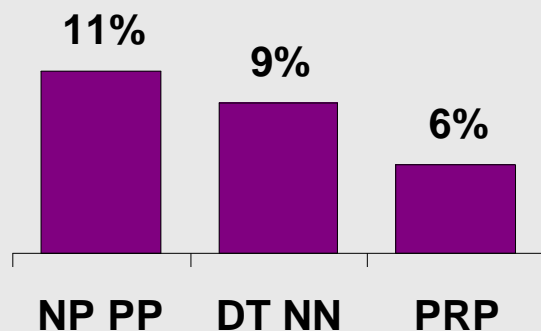
- $p(\text{VP} \rightarrow \text{V NP NP}) = 0.00151$
- $p(\text{VP} \rightarrow \text{V NP NP} \mid \text{said}) = 0.00001$
- $p(\text{VP} \rightarrow \text{V NP NP} \mid \text{gave}) = 0.01980$



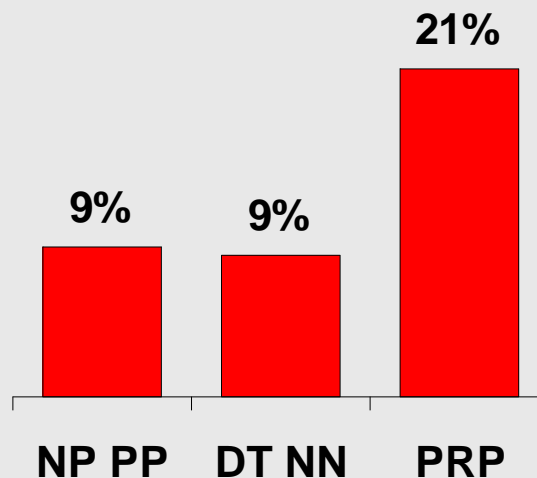
Учет контекста

- Предположение о независимости PCFG часто слишком сильное

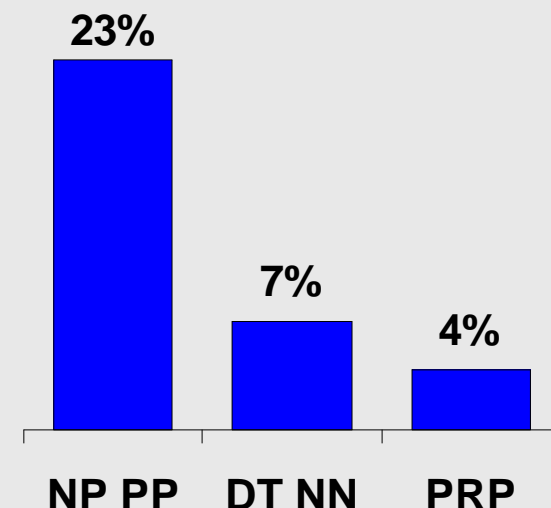
все NPs



NPs под S

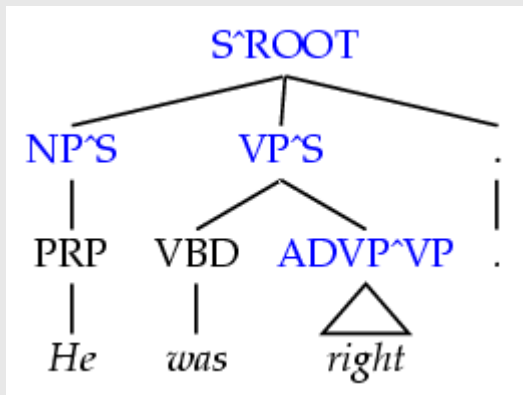


NPs под VP

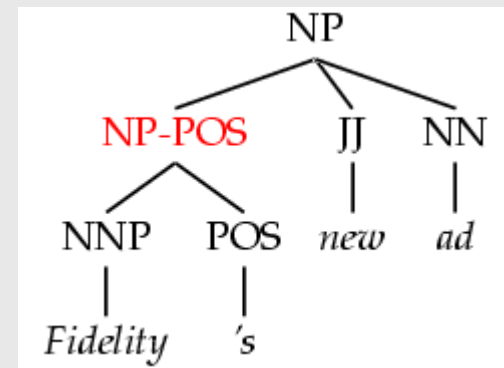


Переопределение тегов

+ родительский тег
[Johnson 98]



Посессивные NPs



Набор тегов на основе обучения

[Petrov and Klein 2006, 2007]

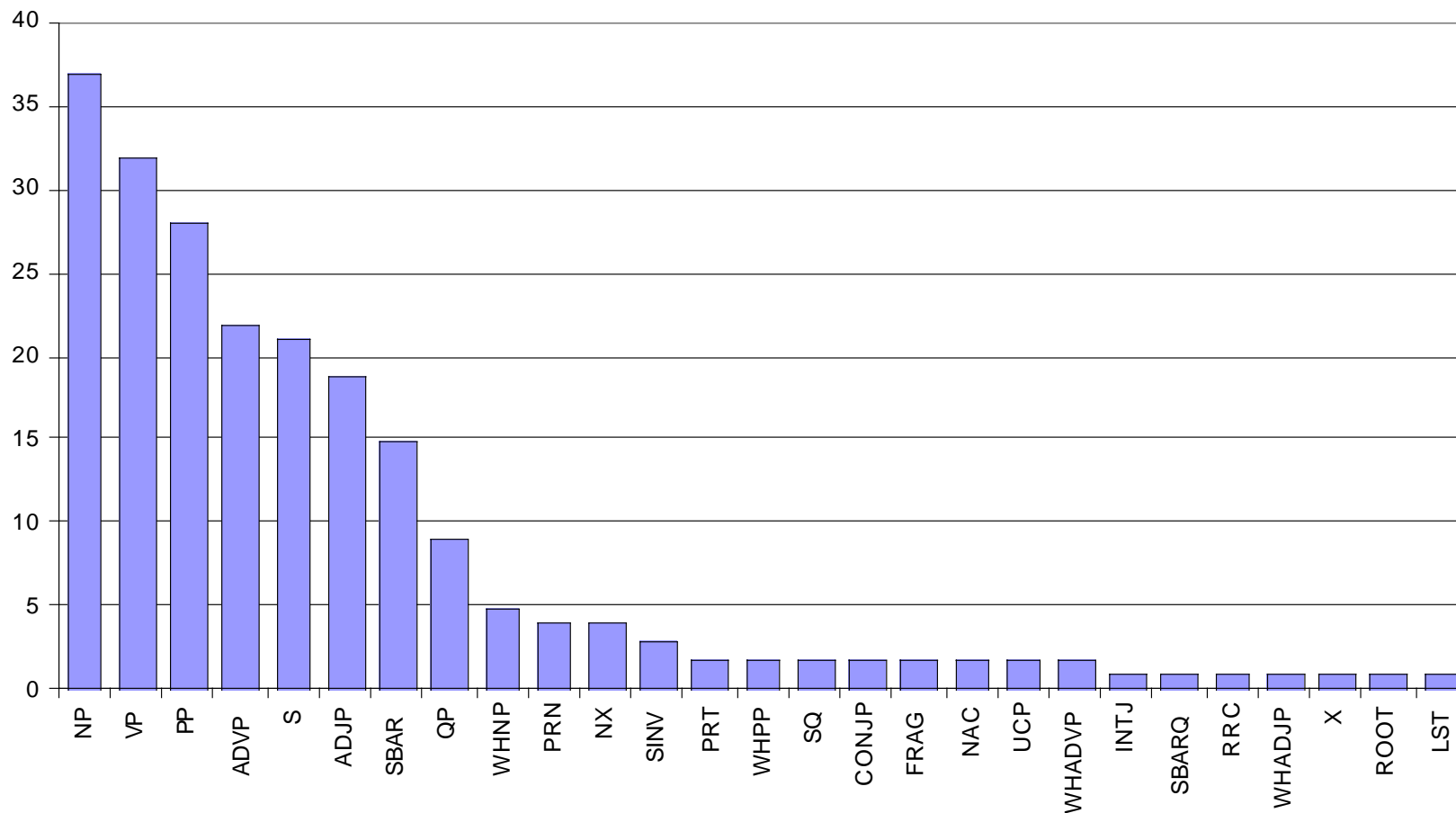
- Proper Nouns (NNP):

NNP-14	Oct.	Nov.	Sept.
NNP-12	John	Robert	James
NNP-2	J.	E.	L.
NNP-1	Bush	Noriega	Peters
NNP-15	New	San	Wall
NNP-3	York	Francisco	Street

- Personal pronouns (PRP):

PRP-0	It	He	I
PRP-1	it	he	they
PRP-2	it	them	him

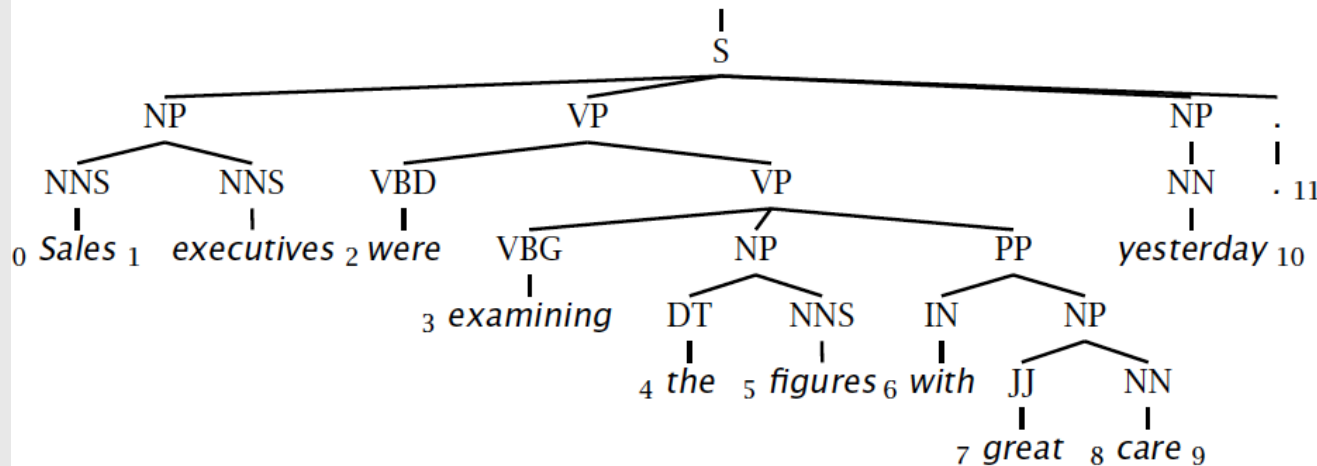
Распределение подкатегорий



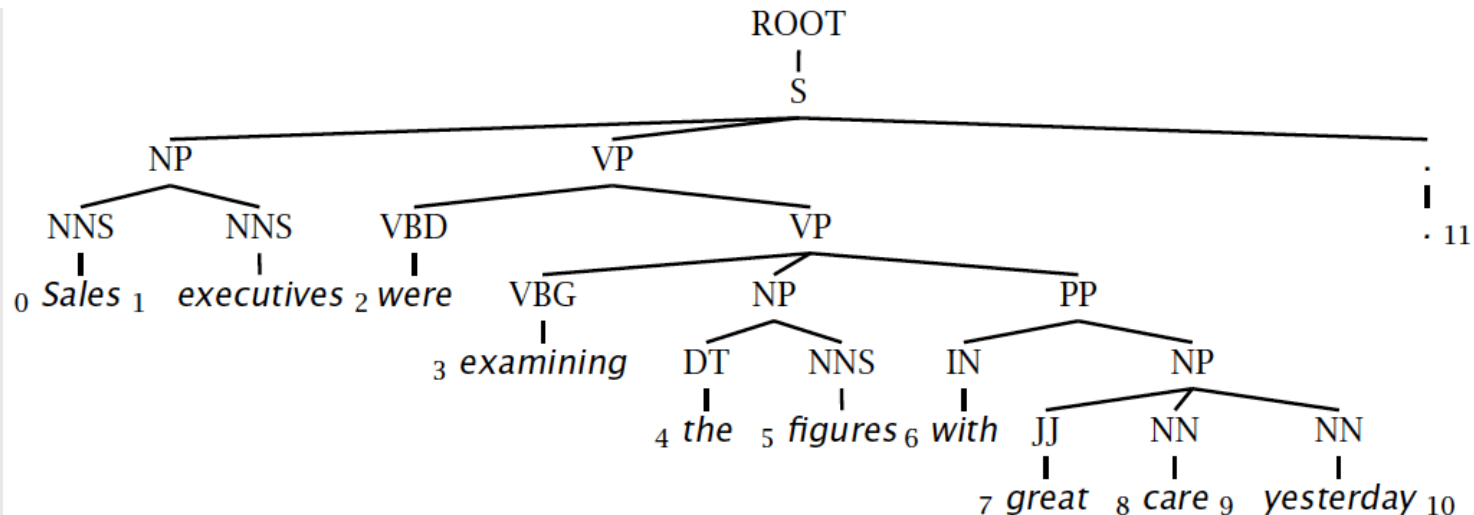
ОЦЕНКА

Оценка парсеров

Gold standard brackets: S-(0:11), NP-(0:2), VP-(2:9), VP-(3:9), NP-(4:6), PP-(6:9), NP-(7,9), NP-(9:10)



Candidate brackets: S-(0:11), NP-(0:2), VP-(2:10), VP-(3:10), NP-(4:6), PP-(6:10), NP-(7,10)



Оценка парсеров – 2

Золотой стандарт:

S-(0:11), NP-(0:2), VP-(2:9), VP-(3:9), NP-(4:6), PP-(6-9), NP-(7,9), NP-(9:10)

Оцениваемый разбор:

S-(0:11), NP-(0:2), VP-(2:10), VP-(3:10), NP-(4:6), PP-(6-10), NP-(7,10)

Labeled Precision	$3/7 = 42.9\%$
Labeled Recall	$3/8 = 37.5\%$
LP/LR F1	40.0%
Tagging Accuracy	$11/11 = 100.0\%$

Прогресс

	<i>F1</i> <i>≤ 40 words</i>	<i>F1</i> <i>all words</i>
<i>Parser</i>		
Klein & Manning unlexicalized 2003	86.3	85.7
Matsuzaki et al. simple EM latent states 2005	86.7	86.1
Charniak generative, lexicalized (“maxent inspired”) 2000	90.1	89.5
Petrov and Klein NAACL 2007	90.6	90.1
Charniak & Johnson discriminative reranker 2005	92.0	91.4
Fossum & Knight 2009 combining constituent parsers		92.4

ПАРСЕРЫ ЗАВИСИМОСТЕЙ

MaltParser

[Nivre et al. 2008]

- жадный (arc-eager) дискриминативный парсер зависимостей
- устанавливает отношение слева направо, не дожидаясь нахождения всех следующих зависимостей
- на каждом этапе – решение об установлении отношения (и его типа)
- выбор действия на основе классификатора, обучается на размеченных данных
- Компоненты:
 - стек σ (на старте содержит символ ROOT)
 - буфер β (на старте содержит слова предложения)
 - множество отношений A (на старте пусто)
 - набор действий

Действия

Start: $\sigma = [\text{ROOT}]$, $\beta = w_1, \dots, w_n$, $A = \emptyset$

1. Left-Arc_r $\sigma | w_i, w_j | \beta, A \Rightarrow \sigma, w_j | \beta, A \cup \{r(w_j, w_i)\}$

Precondition: $r'(w_k, w_i) \notin A$, $w_i \neq \text{ROOT}$

2. Right-Arc_r $\sigma | w_i, w_j | \beta, A \Rightarrow \sigma | w_i | w_j, \beta, A \cup \{r(w_i, w_j)\}$

3. Reduce $\sigma | w_i, \beta, A \Rightarrow \sigma, \beta, A$

Precondition: $r'(w_k, w_i) \in A$

4. Shift $\sigma, w_i | \beta, A \Rightarrow \sigma | w_i, \beta, A$

Finish: $\beta = \emptyset$

Пример

1. Left-Arc_r $\sigma | w_i, w_j | \beta, A \rightarrow \sigma, w_j | \beta, AU\{r(w_i, w_j)\}$
Precondition: $(w_k, r', w_i) \notin A, w_i \neq \text{ROOT}$
2. Right-Arc_r $\sigma | w_i, w_j | \beta, A \rightarrow \sigma | w_i | w_j, \beta, AU\{r(w_i, w_j)\}$
3. Reduce $\sigma | w_i, \beta, A \rightarrow \sigma, \beta, A$
Precondition: $(w_k, r', w_i) \in A$
4. Shift $\sigma, w_i | \beta, A \rightarrow \sigma | w_i, \beta, A$

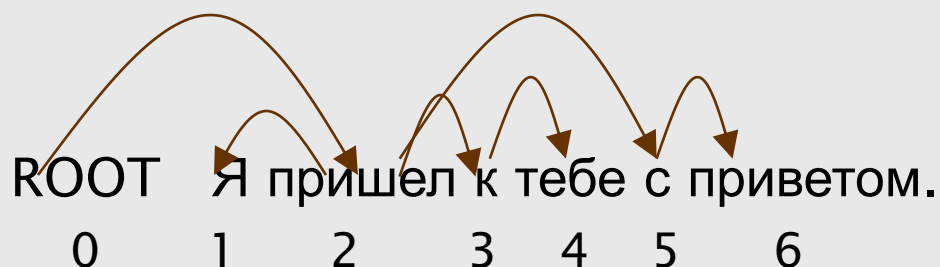
Я пришел к тебе с приветом.

	σ	β	A
	[ROOT]	[я, пришел, к, тебе ...]	\emptyset
Shift	[ROOT, я]	[пришел, к, ...]	\emptyset
LA _{nsubj}	[ROOT]	[пришел, к, ...]	{nsubj (пришел, я)} = A ₁
RA _{root}	[ROOT, пришел]	[к, тебе, ...]	A ₁ U {root(ROOT, пришел)} = A ₂
RA _{prep}	[ROOT, пришел, к]	[тебе, с, ...]	A ₂ U {prep(пришел, к)} = A ₃
RA _{pobj}	[ROOT, пришел, к, тебе]	[с, приветом, .]	A ₃ U {pobj(к, тебе)} = A ₄
Reduce	[ROOT, пришел, к]	[с, приветом, .]	A ₄
Reduce	[ROOT, пришел]	[с, приветом, .]	A ₄
RA _{prep}	[ROOT, пришел, с]	[приветом, .]	A ₄ U {prep(пришел, с)} = A ₅
RA _{pobj}	[ROOT, пришел, с, приветом]	[.]	A ₅ U {pobj(с, приветом)} = A ₆
Reduce	[ROOT, пришел, с]	[.]	A ₆
Reduce	[ROOT, пришел]	[.]	A ₆
RA _{punc}	[ROOT, пришел, .]	[]	A ₆ U {(punc(пришел, .))} = A ₇

Как выбрать действие?

- Обучаем классификатор (например, SVM) на основе размеченных данных
 - нетипизированные отношения: 4 «класса»
 - типизированные отношения: $|R| * 2 + 2$
- Признаки:
 - слово на вершине стека/первое слово в буфере,
 - части речи,
 - расстояние,
 - слова между,
 - валентность и т.д.
- нет поиска вариантов → производительность (!)
- качество сравнимо с LPCFGs

Оценка



$$\text{Acc} = \frac{\text{\# correct deps}}{\text{\# of deps}}$$

$$\text{UAS} = 5 / 6 = 83\%$$

$$\text{LAS} = 4 / 6 = 67\%$$

Золотой стандарт

1	2	Я	nsubj
2	0	пришел	root
3	2	к	prep
4	3	тебе	pobj
5	2	с	prep
6	5	приветом	pobj

Разбор

1	2	Я	nsubj
2	0	пришел	root
3	2	к	prep
4	3	тебе	dobj
5	1	с	prep
6	5	приветом	pobj

ПАРСЕРЫ ДЛЯ РУССКОГО

AOT

Input Your text: Russian ▾

Лондонский футбольный клуб «Челси» объявил о назначении итальянца Антонио Конте новым главным тренером.

Submit Request

гл_личн

прил_сущ

пг

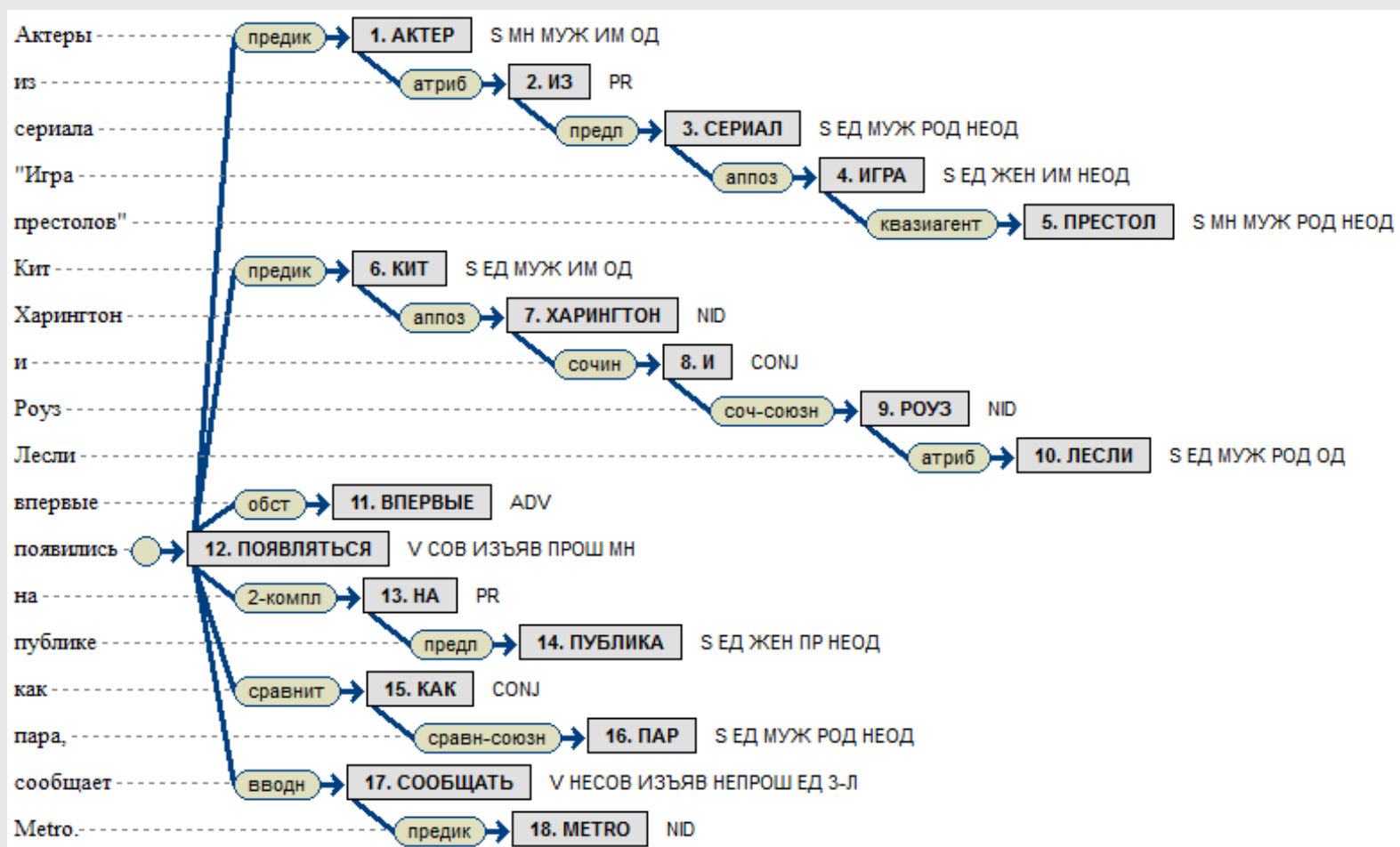
генит_иг

Лондонский футбольный клуб « Челси » объявил о назначении

прил_сущ

итальянца Антонио Конте новым главным тренером .

ЭТАП-3



Томи́та-парсер

- Реализует контекстно-свободные грамматики
- Алгоритм GLR – парсинга
(<http://ru.wikipedia.org/wiki/GLR-парсер>)
- Не используется для полного синтаксического разбора предложения
- Используется для выделения именованных сущностей и отношений (фактов)

MaltParser

Table 4. Parsing results on development set of SynTagRus; labeled attachment score (LAS) and unlabeled attachment score (UAS)

	LAS	UAS
SynTagRus tags, poly-SVM	83.4	89.4
MTE tags, poly-SVM	82.8	88.8
MTE tags, linear SVM	82.2	88.0

<http://corpus.leeds.ac.uk/mocky/>

[Sharoff& Nivre, 2011]

SyntaxNet

Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source

Thursday, May 12, 2016

Posted by Slav Petrov, Senior Staff Research Scientist

At Google, we spend a lot of time thinking about how [computer systems](#) can [read](#) and [understand human language](#) in order [to process it](#) in [intelligent ways](#). Today, we are excited to share the fruits of our research with the broader community by releasing [SyntaxNet](#), an open-source neural network framework implemented in [TensorFlow](#) that provides a foundation for [Natural Language Understanding](#) (NLU) systems. Our release includes all the code needed to train new SyntaxNet models on your own data, as well as *Parsey McParseface*, an English parser that we have trained for you and that you can use to analyze English text.

Parsey McParseface is built on powerful machine learning algorithms that learn to analyze the linguistic structure of language, and that can explain the functional role of each word in a given sentence. Because Parsey McParseface is the [most accurate such model in the world](#), we hope that it will be useful to developers and researchers interested in automatic extraction of information, translation, and other core applications of NLU.

How does SyntaxNet work?

SyntaxNet is a framework for what's known in academic circles as a [syntactic parser](#), which is a key first component in many NLU systems. Given a sentence as input, it tags each word with a part-of-speech (POS) tag that describes the word's syntactic function, and it determines the syntactic relationships between words in the sentence, represented in the dependency parse tree. These syntactic relationships are directly related to the underlying meaning of the sentence in question. To take a very simple example, consider the following dependency tree for *Alice saw Bob*:

<https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html>

Language	No. tokens	POS	fPOS	Morph	UAS	LAS
Russian-SynTagRus	107737	98.27%	-	94.91%	91.68%	87.44%
Russian	9573	95.27%	95.02%	87.75%	81.75%	77.71%

CoNLL-U формат

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1-2	Vámonos	—	—	—	—	—	—	—	—
1	Vamos	ir	VERB	—	Mood=Imp Number=Plur Person=1	0	root	—	—
2	nos	nosotros	PRON	—	PronType=Per Number=Plur Person=1	1	expl	—	—
3-4	al	—	—	—	—	—	—	—	—
3	a	a	ADP	—	—	5	case	—	—
4	el	el	DET	—	Definite=Def Number=Sing —	5	det	—	—
5	mar	mar	NOUN	—	Number=Sing Gender=Masc	1	nmod	—	—
6	.	.	.	—	—	1	punct	—	—

[Nivre]

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for tokens with multiple words.
2. FORM: Word form or punctuation symbol.
3. LEMMA: Lemma or stem of word form.
4. UPOSTAG: [Universal part-of-speech tag](#) drawn from our revised version of the Google universal POS tags.
5. XPOSTAG: Language-specific part-of-speech tag; underscore if not available.
6. FEATS: List of morphological features from the [universal feature inventory](#) or from a defined [language-specific extension](#); underscore if not available.
7. HEAD: Head of the current token, which is either a value of ID or zero (0).
8. DEPREL: [Universal dependency relation](#) to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
9. DEPS: List of secondary dependencies (head-deprel pairs).
10. MISC: Any other annotation.

<http://universaldependencies.org/format.html>

Примеры

```

1  как _   ADV _   Degree=Pos | fPOS=ADV++ 2   advmod _ _
2  сделать _ VERB _   Aspect=Perf | VerbForm=Inf | fPOS=VERB++ 0   ROOT _ _
3  из _   ADP _   fPOS=ADP++ 4   case _ _
4  бумаги _ NOUN _   Animacy=Inan | Case=Gen | Gender=Fem | Number=Sing | fPOS=NOUN++ 2   dobj _ _
5  розу _   NOUN _   Animacy=Inan | Case=Dat | Gender=Masc | Number=Sing | fPOS=NOUN++ 2   nmod _ _

```

```

1  как _   ADV _   Degree=Pos | fPOS=ADV++ 2   advmod _ _
2  сделать _ VERB _   Aspect=Perf | VerbForm=Inf | fPOS=VERB++ 0   ROOT _ _
3  мыльные _ ADJ _   Animacy=Inan | Case=Acc | Degree=Pos | Number=Plur | fPOS=ADJ++ 4   amod _ _
4  пузыри _ NOUN _   Animacy=Inan | Case=Acc | Gender=Masc | Number=Plur | fPOS=NOUN++ 2   dobj _ _
5  в _   ADP _   fPOS=ADP++ 7   case _ _
6  домашних _ ADJ _   Case=Loc | Degree=Pos | Number=Plur | fPOS=ADJ++ 7   amod _ _
7  условиях _ NOUN _   Animacy=Inan | Case=Loc | Gender=Neut | Number=Plur | fPOS=NOUN++ 2   nmod _ _

```

```

1  как _   ADV _   fPOS=SCONJ++ 2   advmod _ _
2  быстро _ ADV _   Degree=Pos | fPOS=ADV++ 3   advmod _ _
3  похудеть _ VERB _   Aspect=Perf | VerbForm=Inf | fPOS=VERB++ 0   ROOT _ _
4  на _   ADP _   fPOS=ADP++ 6   case _ _
5  10 _   NUM _   fPOS=NUM++ 6   nummod _ _
6  кг _   NOUN _   Animacy=Inan | Case=Gen | Gender=Masc | Number=Plur | fPOS=NOUN++ 3   dobj _ _

```

ОЦЕНКА

Оценка парсеров (2011-2012)

- Коллекция:
 - смесь жанров (НКРЯ)
 - новости (РОМИП)
- Токенизация
- Золотой стандарт – 800 предложений (500+300)
- метрика: доля правильных зависимостей (+ ручная проверка несовпадений)
- 7 участников
- сложности унификации

Результаты (UAS)

	P	R	F1	
Trieste	0,952	0,983	0,967	Compreno
Marceille	0,933	0,981	0,956	ЭТАП–3
Barcelona	0,895	0,980	0,935	SyntAutom
Toulon	0,889	0,947	0,917	SemSyn
Brega	0,863	0,980	0,917	Dictum
Nice	0,856	0,860	0,858	Semantic analyzer group
Napoli	0,789	0,975	0,872	
				AotSoft