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, phrase 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

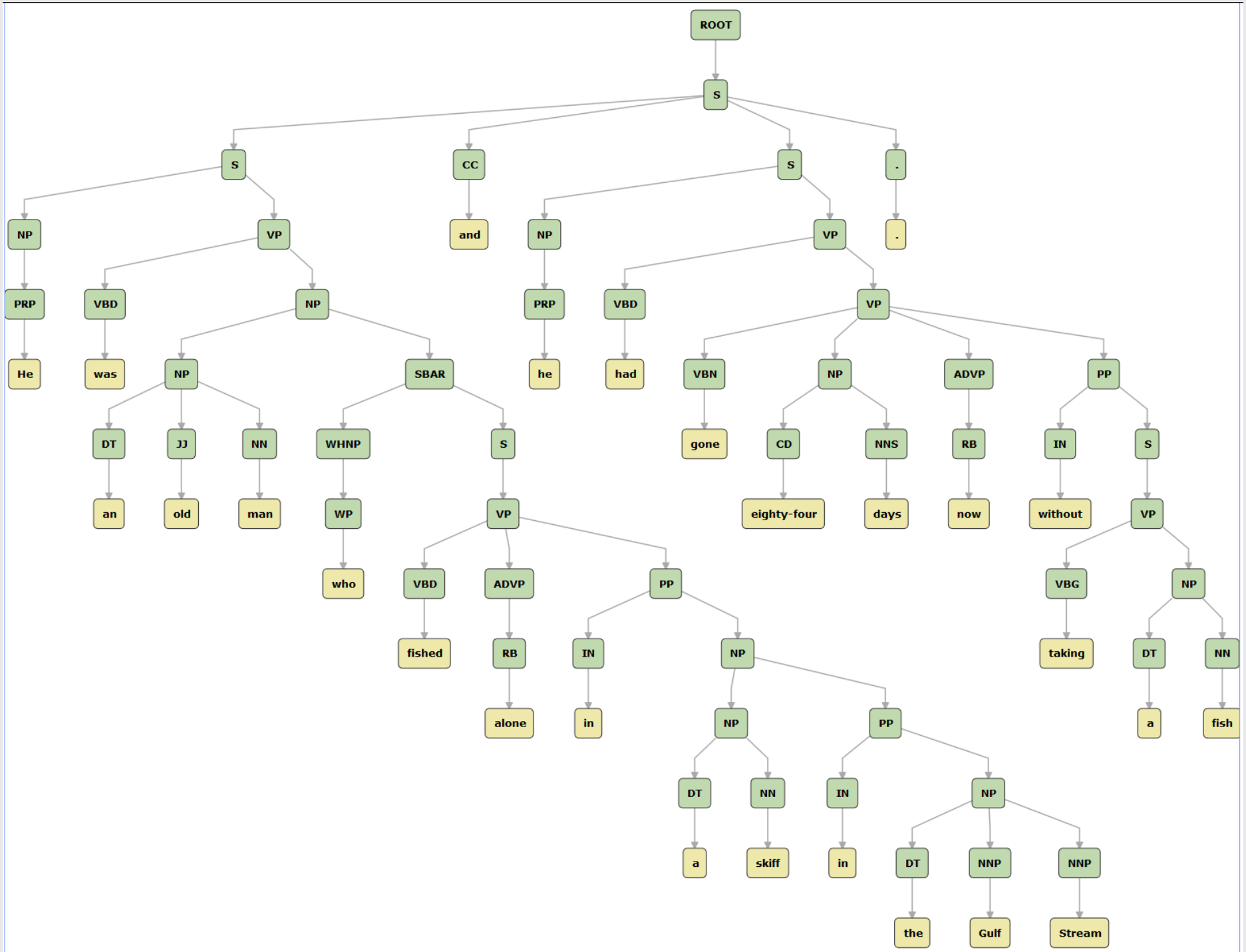
Phrase Level

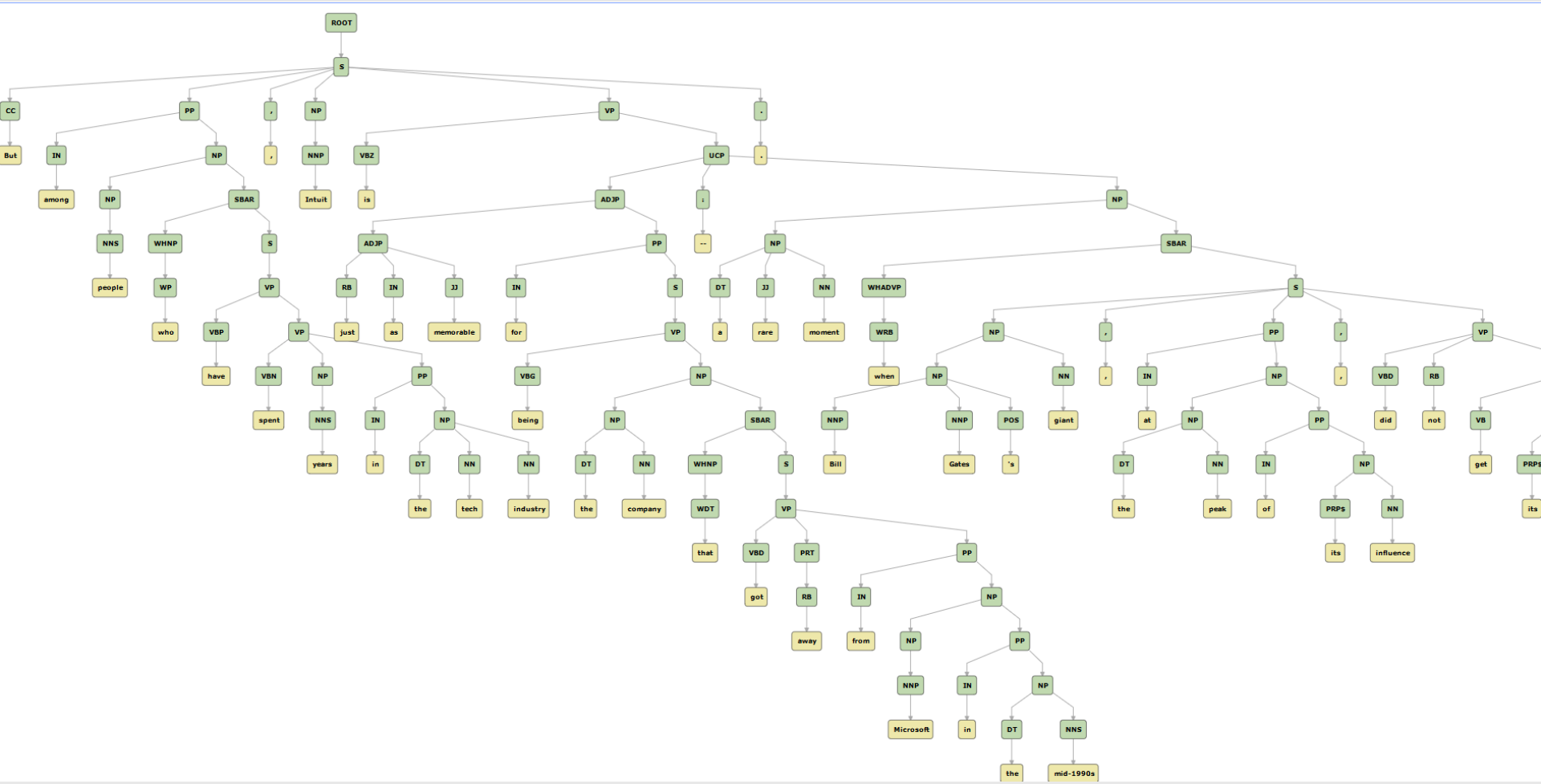
ADJP - Adjective Phrase.
ADVP - Adverb Phrase.
CONJP - Conjunction Phrase.
FRAG - Fragment.
INTJ - Interjection. Corresponds approximately to the part-of-speech tag UH.
LST - List marker. Includes surrounding punctuation.
NAC - Not a Constituent; used to show the scope of certain prenominal modifiers 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 - Vereb 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.

http://web.mit.edu/6.863/www/PennTreebankTags.html

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


ROOT
S
S
CC
S
.
NP
VP
and
NP
VP
.
PRP
VBD
NP
PRP
VBD
VP
He
was
NP
SBAR
he
had
VBN
NP
ADVP
PP
DT
JJ
NN
WHNP
S
gone
CD
NNS
RB
IN
S
an
old
man
WP
VP
eighty-four
days
now
without
VP
who
VBD
ADVP
PP
VBG
NP
fished
RB
IN
NP
taking
DT
NN
alone
in
NP
PP
a
fish
DT
NN
IN
NP
a
skiff
in
DT
NNP
NNP
the
Gulf
Stream


ROOT
S
CC
But
PP
IN
among
NP
NNS
people
SBAR
WHNP
WP
who
S
VP
VBP
have
VP
VBN
spent
NP
NNS
years
PP
IN
in
NP
DT
the
NN
tech
NN
industry
,
,
NP
NNP
Intuit
VP
VBZ
is
ADJP
ADJP
RB
just
IN
as
JJ
memorable
PP
IN
for
S
VP
VBG
being
NP
NP
DT
the
NN
company
SBAR
WHNP
WDT
that
S
VP
VBD
got
PRT
RB
away
PP
IN
from
NP
NNP
Microsoft
PP
IN
in
NP
DT
the
NNS
mid-1990s
UCP
;
--
.
.
NP
DT
a
JJ
rare
NN
moment
NP
WHADVP
WRB
when
NP
NNP
Bill
NNP
Gates
POS
's
NN
giant
SBAR
S
NP
,
,
WHADVP
WRB
when
NP
NNP
Bill
NNP
Gates
POS
's
NN
giant
,
,
PP
IN
at
NP
DT
the
NN
peak
PP
IN
of
NP
PRP$
its
NN
influence
,
,
VP
VBD
did
RB
not
VB
get
PRP$
its

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


root
В Москве, на Патриарших прудах, появились два гражданина.

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 —

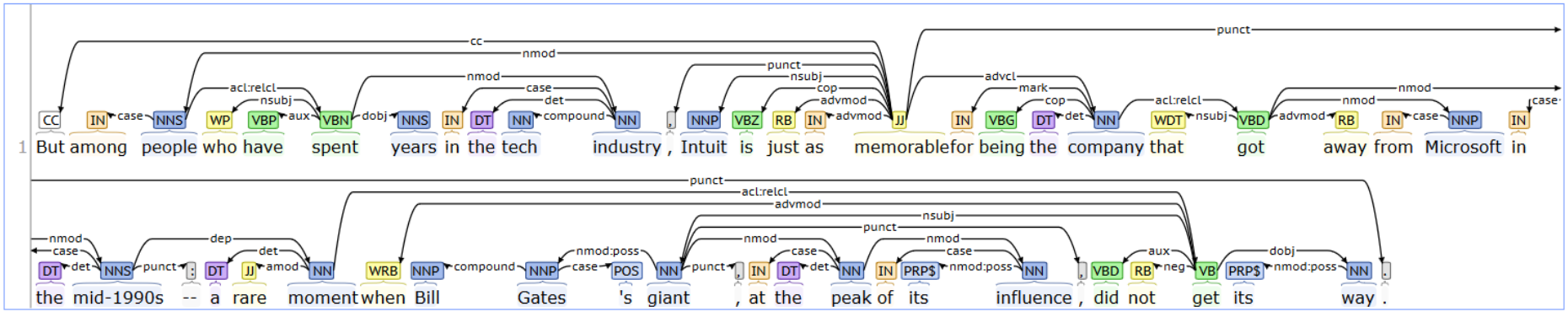
dependency parse ×

— 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*

цит. по [Тестелец]

UD: типы зависимостей

Core dependents of clausal predicates

Nominal dep	*Predicate dep*	
nsubj	csubj	
nsubjpass	csubjpass	
dobj	ccomp	xcomp
iobj		

Non-core dependents of clausal predicates

Nominal dep	*Predicate dep*	*Modifier word*
nmod	advcl	advmod
		neg

Special clausal dependents

Nominal dep	*Auxiliary*	*Other*
vocative	aux	mark
discourse	auxpass	punct
expl	cop	

Noun dependents

Nominal dep	*Predicate dep*	*Modifier word*
nummod	acl	amod
appos		det
nmod		neg

Compounding and unanalyzed

compound	mwe	goeswith
name	foreign	

Coordination

conj	cc	punct

Case-marking, prepositions, possessive

case

Loose joining relations

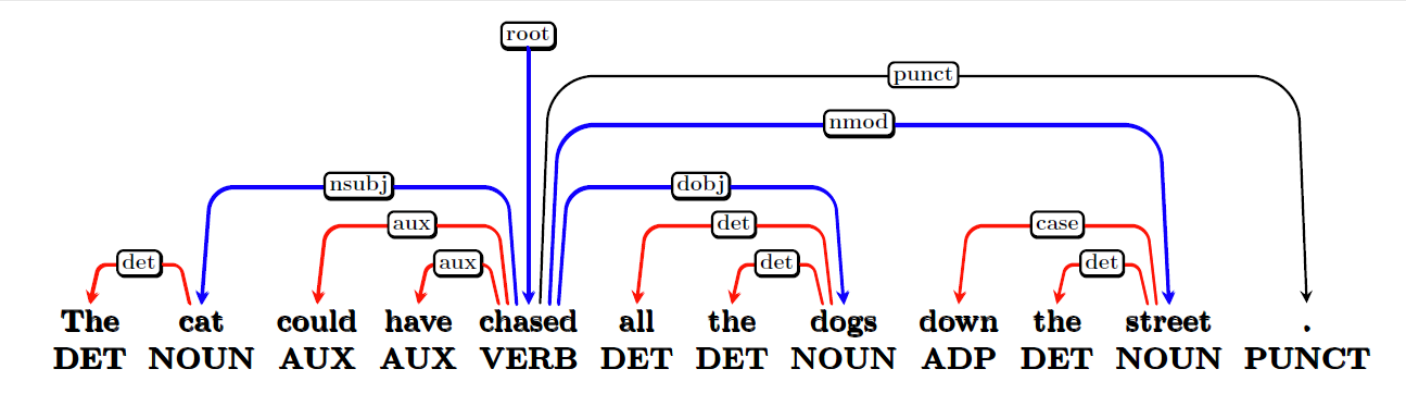
list	parataxis	remnant
dislocated		reparandum

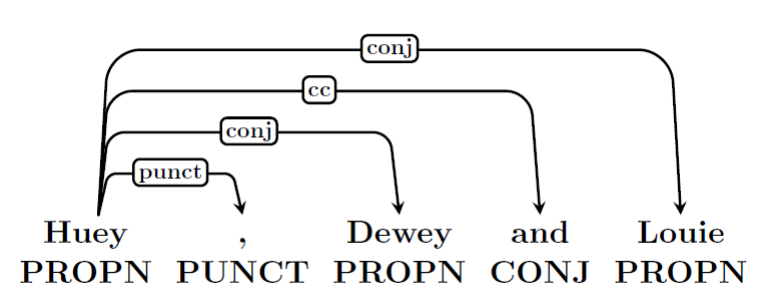
Other

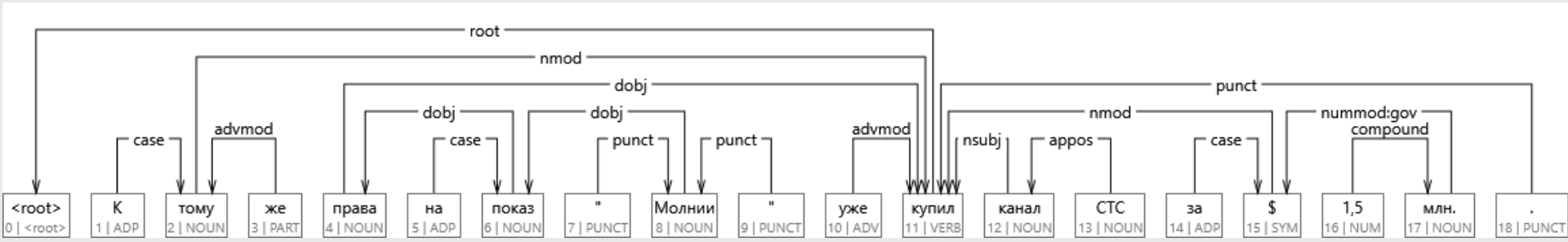
Sentence head	*Unspecified dependency*
root	dep

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

UD: пример


root
punct
nmod
nsubj
dobj
aux
det
case
det
aux
det
det
det
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT


conj
cc
conj
punct
Huey , Dewey and Louie
PROPN PUNCT PROPN CONJ PROPN


root
nmod
dobj
punct
case
advmod
dobj
dobj
case
punct
punct
advmod
nmod
nsubj
appos
case
nummod:gov
compound
<root> 0 | <root>
К 1 | ADP
тому 2 | NOUN
же 3 | PART
права 4 | NOUN
на 5 | ADP
показ 6 | NOUN
" 7 | PUNCT
Молнии 8 | NOUN
" 9 | PUNCT
уже 10 | ADV
купил 11 | VERB
канал 12 | NOUN
СТС 13 | NOUN
за 14 | ADP
$ 15 | SYM
1,5 16 | NUM
млн. 17 | NOUN
. 18 | PUNCT

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)))))))
    (. .)))
```

[Marcus et al. 1993, *Computational Linguistics*]

[Chris Manning]

SynTagRus

- 64К 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 страни

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)**

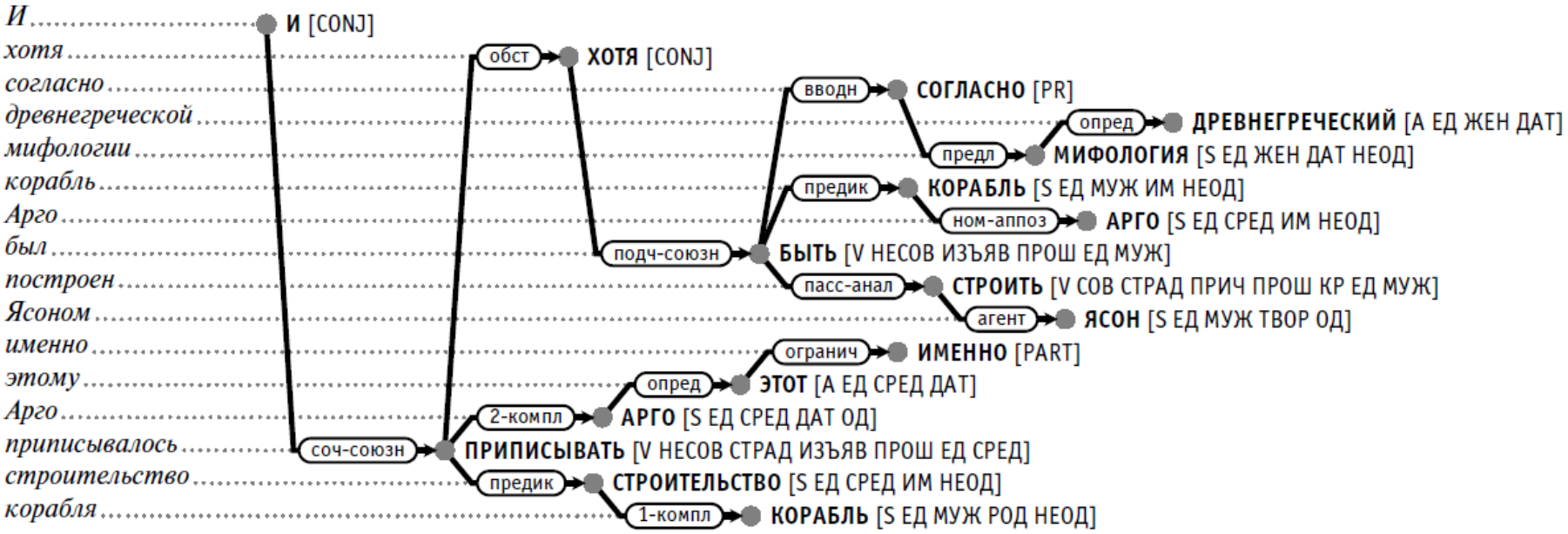
Пятнадцать лет назад, 15 ноября 1988 года, совершил свой полет, закончившийся не повторенной до сих пор автоматической посадкой на посадочную пол
космический **корабль** "Буран". [Советский шаттл] [Показать структуру]
Хотя для первого космонавта наряду с "гагаринским" "Востоком" ОКБ- 256 Павла Цыбина 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 посад
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
В свое время в одной из повестей я написал, что ты искал там **корабль** инопланетян и что Сергей Павлович Королев очень интересовался этой экспедици
- Еще юношей я прочитал статью писателя Казанцева о том, что Тунгусский метеорит - это космический **корабль**, который потерпел катастрофу. [Космонав
Мы использовали отчет Золотова, страниц на пятьсот, и из него следовало, что это был космический **корабль**, который взорвался на высоте пять километ
килотонн. [Космонавт Гречко] [Показать структуру]
У Константина Петровича было какое-то удивительное чутье, он без данных, без расчетов мог скомпоновать **корабль**. [Космонавт Гречко] [Показать структур
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 к себе домой **Все примеры (8)**

SynTagRus – пример



Syntactic treebanks [edit]

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 = (Σ, N, S, R)
  - Σ: терминалы
  - N: нетерминалы
  - S: начальный символ (S ∈ N)
  - R: правила (продукции) в виде $X \to \gamma$
    - $X \in N$; $\gamma \in (N \cup \Sigma)^*$

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

$N \rightarrow$ people
$N \rightarrow$ fish
$N \rightarrow$ tanks
$N \rightarrow$ rods
$V \rightarrow$ people
$V \rightarrow$ fish
$V \rightarrow$ tanks
$P \rightarrow$ with

[Chris Manning]

Нормальная форма Хомского

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

[Chris Manning]

Пример

S → NP VP

VP → V NP

S → V NP

VP → V @VP_V

@VP_V → NP PP

S → V @S_V

@S_V → NP PP

VP → V PP

S → V PP

NP → NP NP

NP → NP PP

NP → P NP

PP → P NP

NP → *people*

NP → *fish*

NP → *tanks*

NP → *rods*

V → *people*

S → *people*

VP → *people*

V → *fish*

S → *fish*

VP → *fish*

V → *tanks*

S → *tanks*

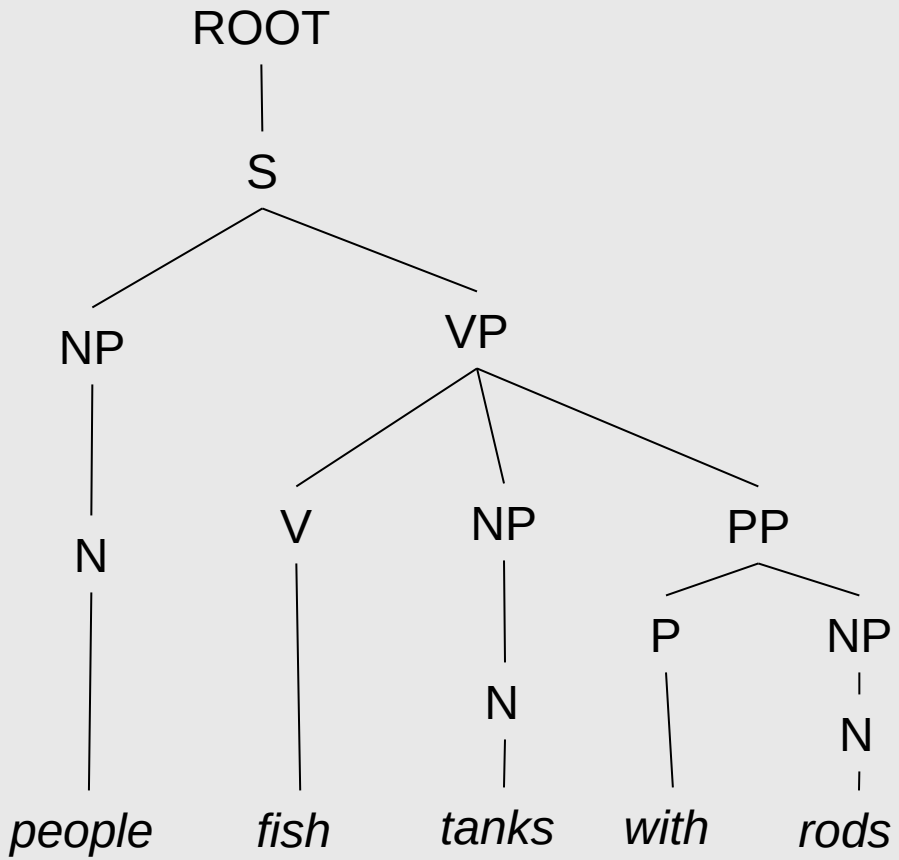
VP → *tanks*

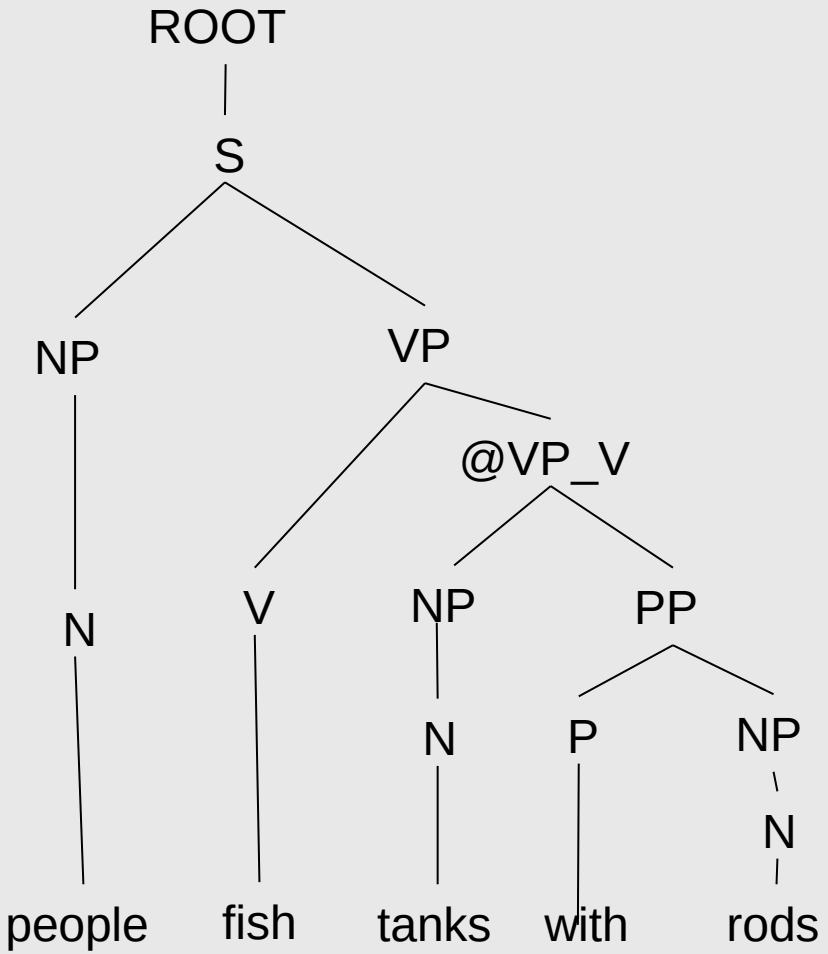
P → *with*

PP → *with*

[Chris Manning]

До и после бинаризации

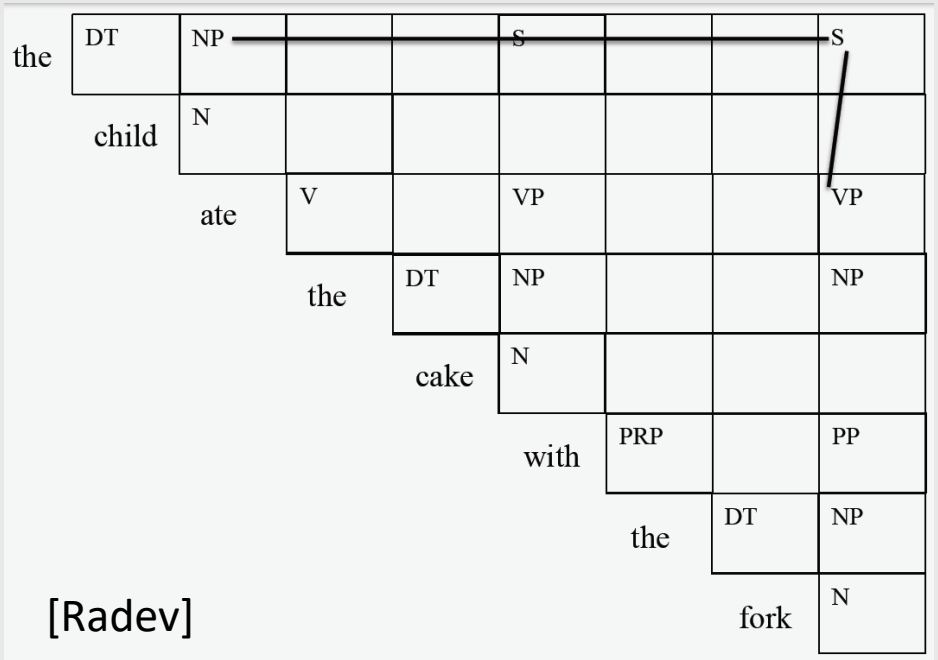




[Chris Manning]

Классические парсеры

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



СТАТИСТИЧЕСКИЕ ПАРСЕРЫ

Вероятностные КСГ

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

$$\forall X \in N, \sum_{X \rightarrow \gamma \in R} P(X \rightarrow \gamma) = 1$$

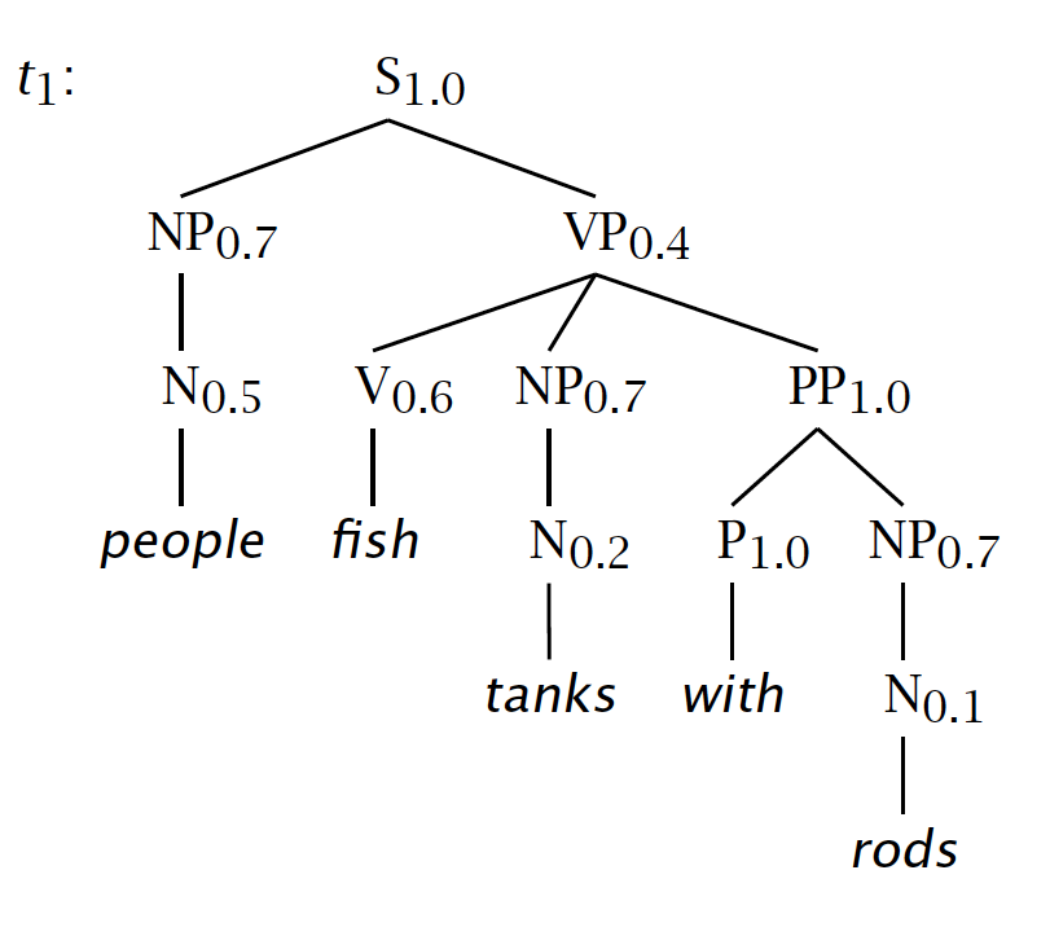
Пример

$S \rightarrow NP\ VP$	1.0
$VP \rightarrow V\ NP$	0.6
$VP \rightarrow V\ NP\ PP$	0.4
$NP \rightarrow NP\ NP$	0.1
$NP \rightarrow NP\ PP$	0.2
$NP \rightarrow N$	0.7
$PP \rightarrow P\ NP$	1.0

$N \rightarrow$ *people*	0.5
$N \rightarrow$ *fish*	0.2
$N \rightarrow$ *tanks*	0.2
$N \rightarrow$ *rods*	0.1
$V \rightarrow$ *people*	0.1
$V \rightarrow$ *fish*	0.6
$V \rightarrow$ *tanks*	0.3
$P \rightarrow$ *with*	1.0

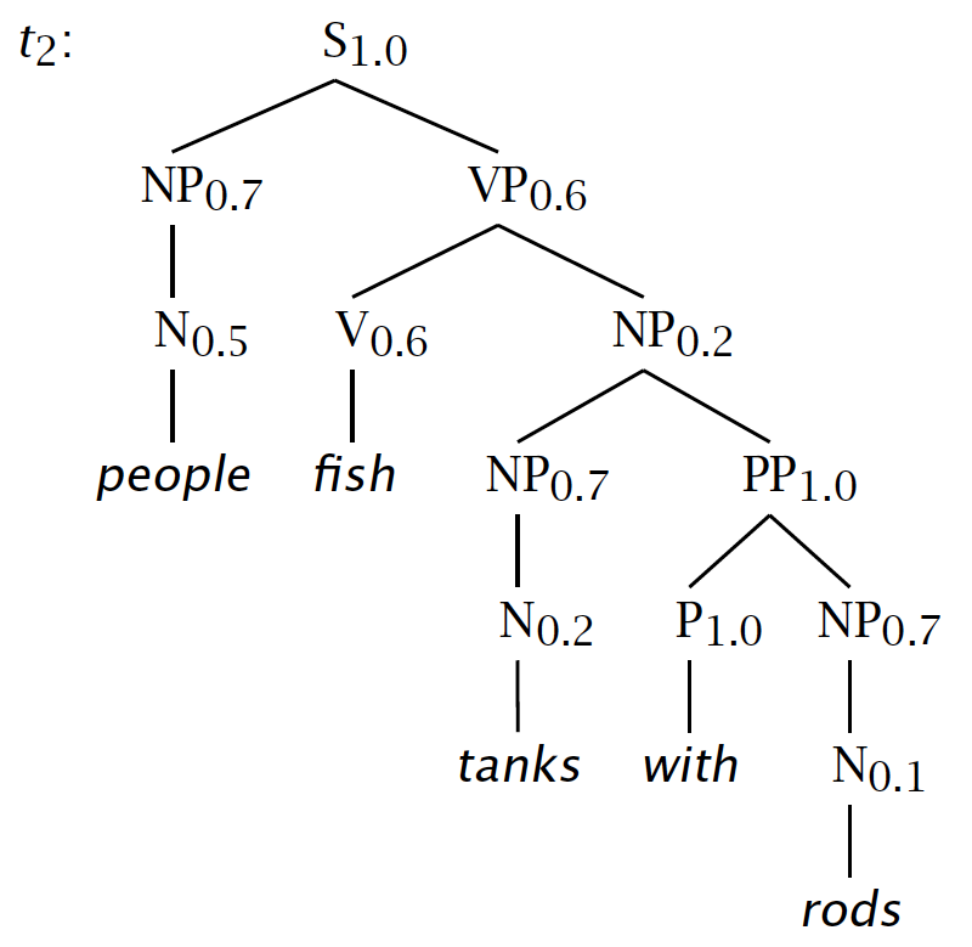
[Chris Manning]

Пример – 1



[Chris Manning]

Пример – 2



[Chris Manning]

Пример: вероятности разбора

- *s* = *people fish tanks with rods*
- $P(t_1)$ = 1.0 × 0.7 × 0.4 × 0.5 × 0.6 × 0.7 × 1.0 × 0.2 × 1.0 × 0.7 × 0.1 = 0.0008232

Verb attach

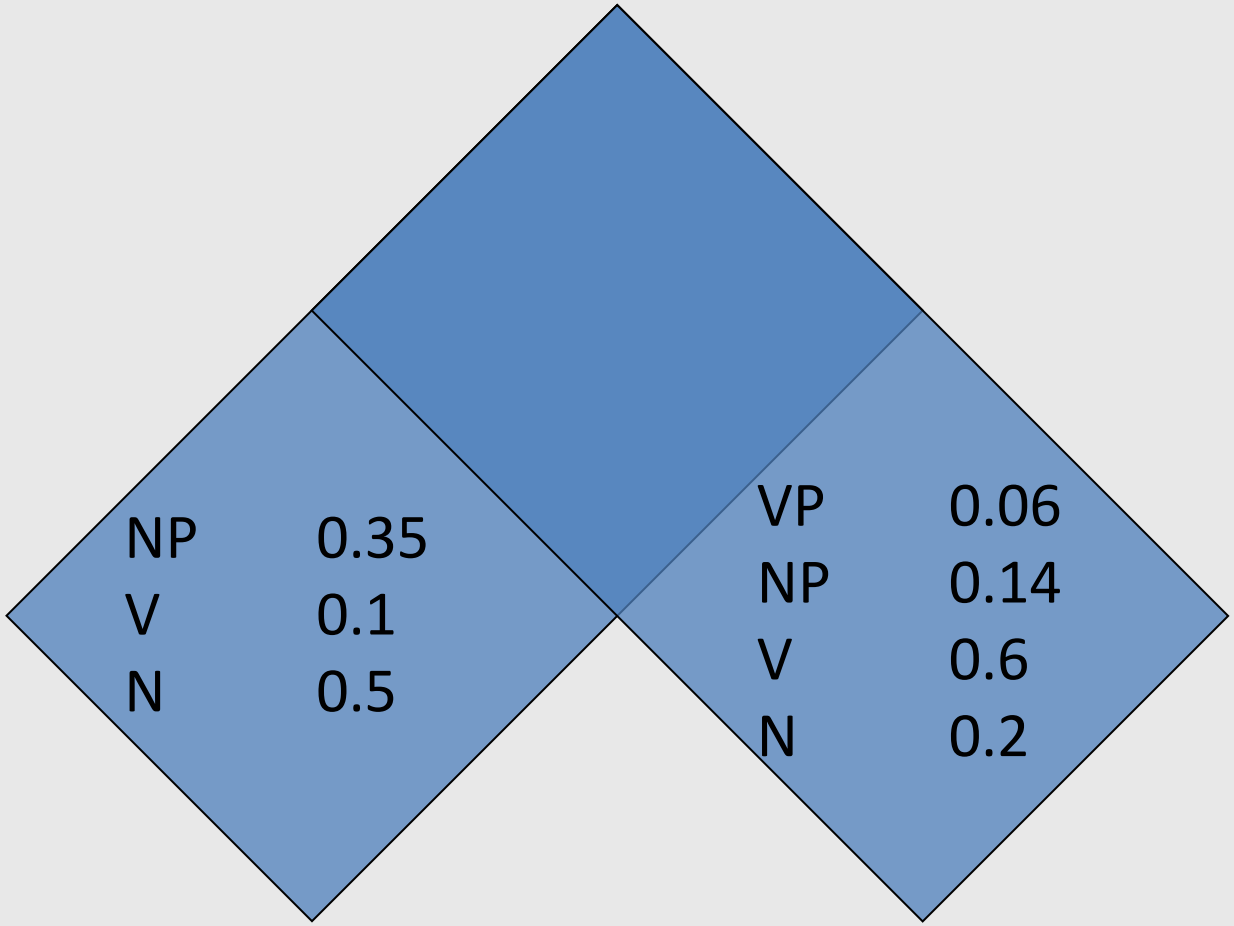
- $P(t_2)$ = 1.0 × 0.7 × 0.6 × 0.5 × 0.6 × 0.2 × 0.7 × 1.0 × 0.2 × 1.0 × 0.7 × 0.1 = 0.00024696

Noun attach

- $P(s) = P(t_1) + P(t_2)$ = 0.0008232 + 0.00024696 = 0.00107016

[Chris Manning]

Viterbi (Max) Scores



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

[Chris Manning]

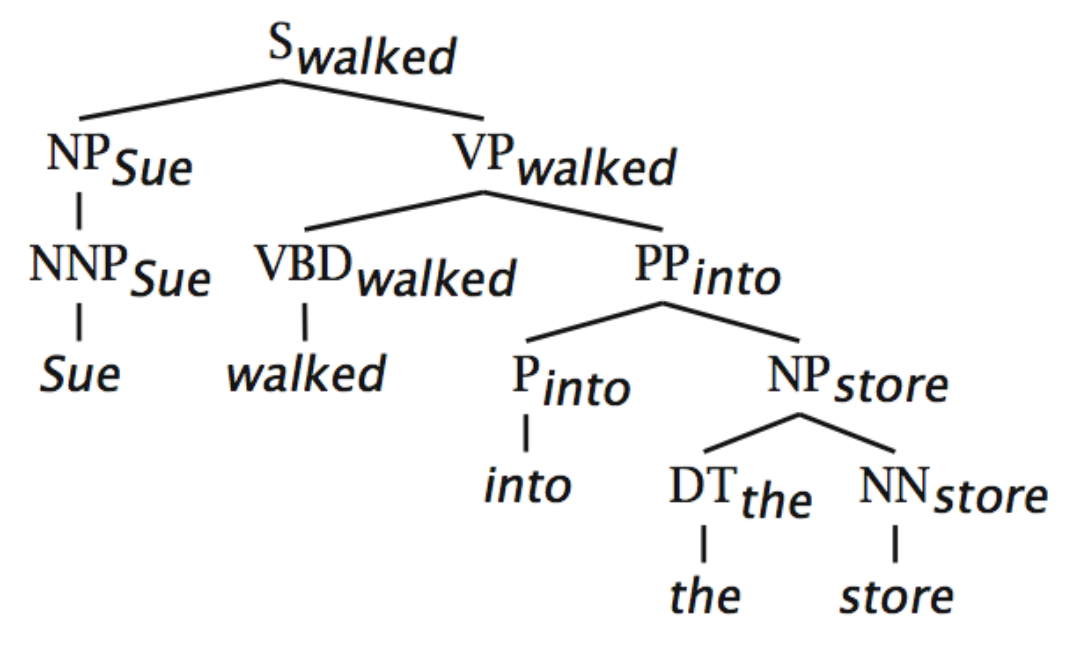
ПОВЫШЕНИЕ КАЧЕСТВА

Как можно улучшить качество?

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

Лексикализация

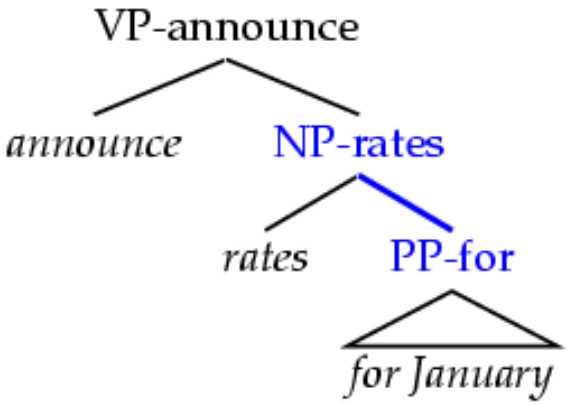
[Magerman 1995, Collins 1997; Charniak 1997]

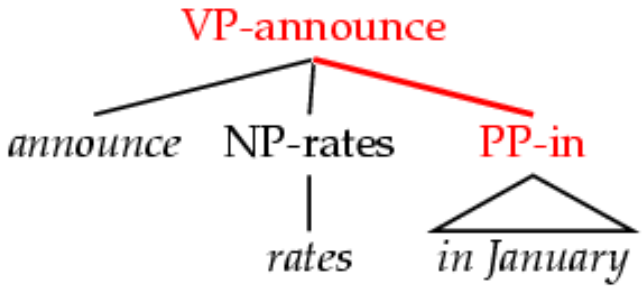


[Chris Manning]

Лексикализация – 2

- p(VP $\rightarrow$ V NP NP) = 0.00151
- p(VP $\rightarrow$ V NP NP | said) = 0.00001
- p(VP $\rightarrow$ V NP NP | gave) = 0.01980

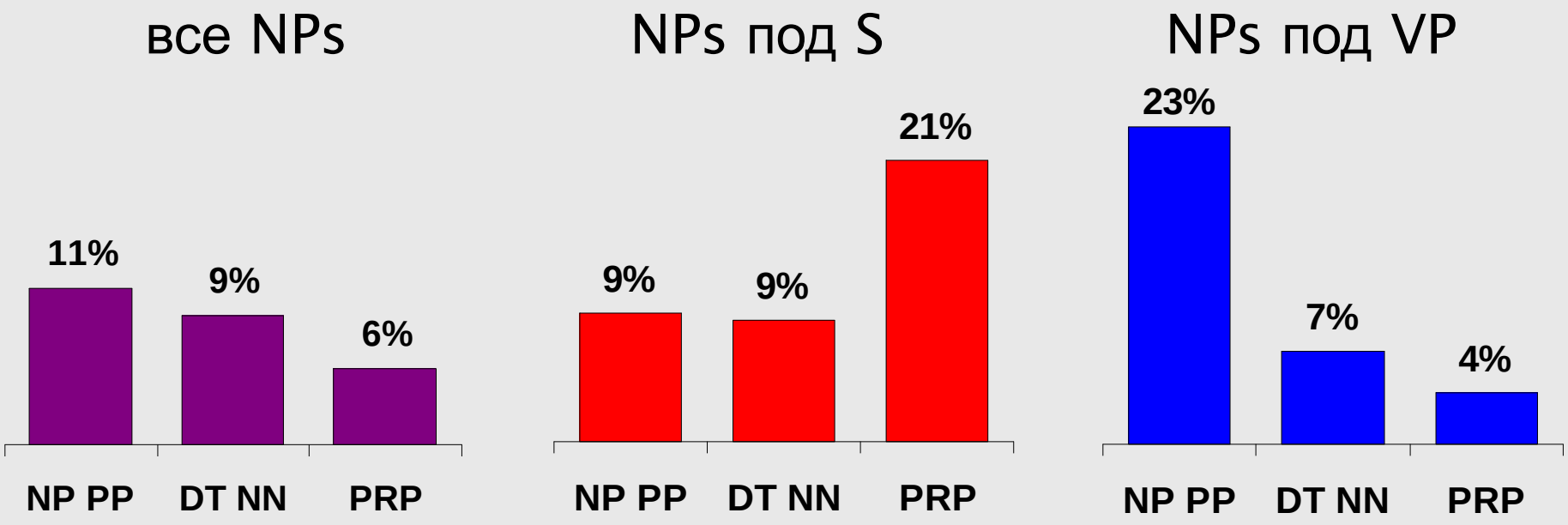




[Chris Manning]

Учет контекста

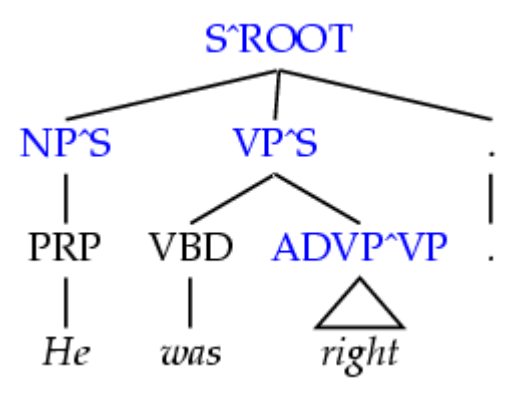
- Предположение о независимости PCFG часто слишком сильное



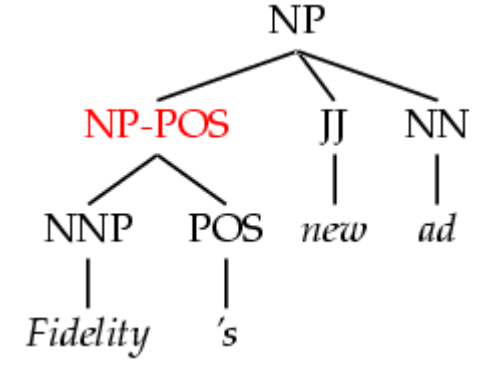
[Chris Manning]

Переопределение тегов

+ родительский тег
[Johnson 98]



Посессивные NPs



[Chris Manning]

Набор тегов на основе обучения

[Petrov and Klein 2006, 2007]

- Proper Nouns (NNP):

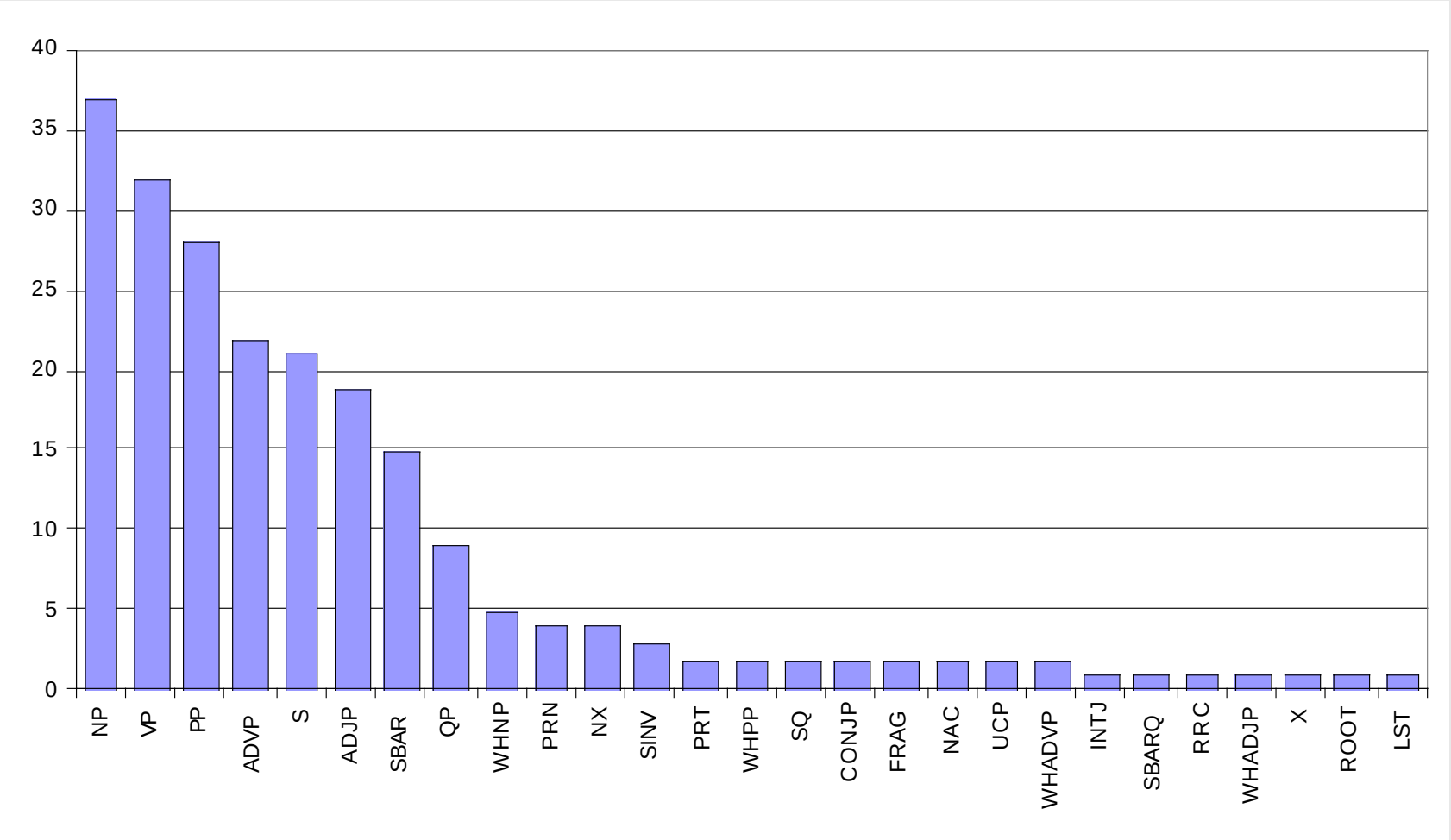
NNP-14	Oct.	Nov.	Sept.
NNP-12	John	Robert	James
NNP-2	J.	E.	L.
NNP-1	Bush	Noriega	Peters
NNP-15	New	San	Wall
NNP-3	York	Francisco	Street

- Personal pronouns (PRP):

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

[Chris Manning]

Распределение подкатегорий

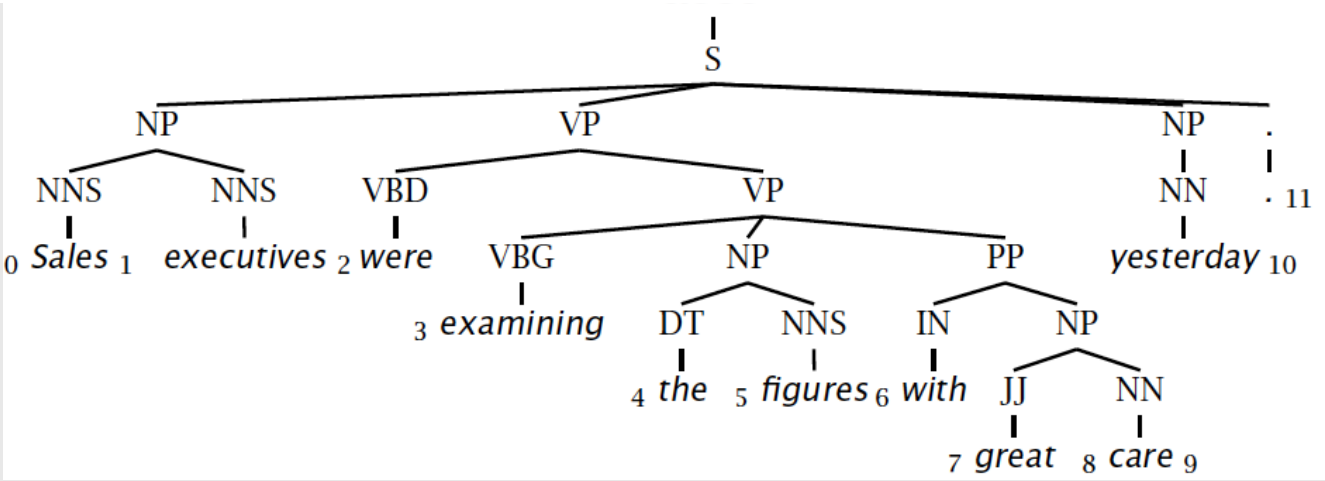


[Chris Manning]

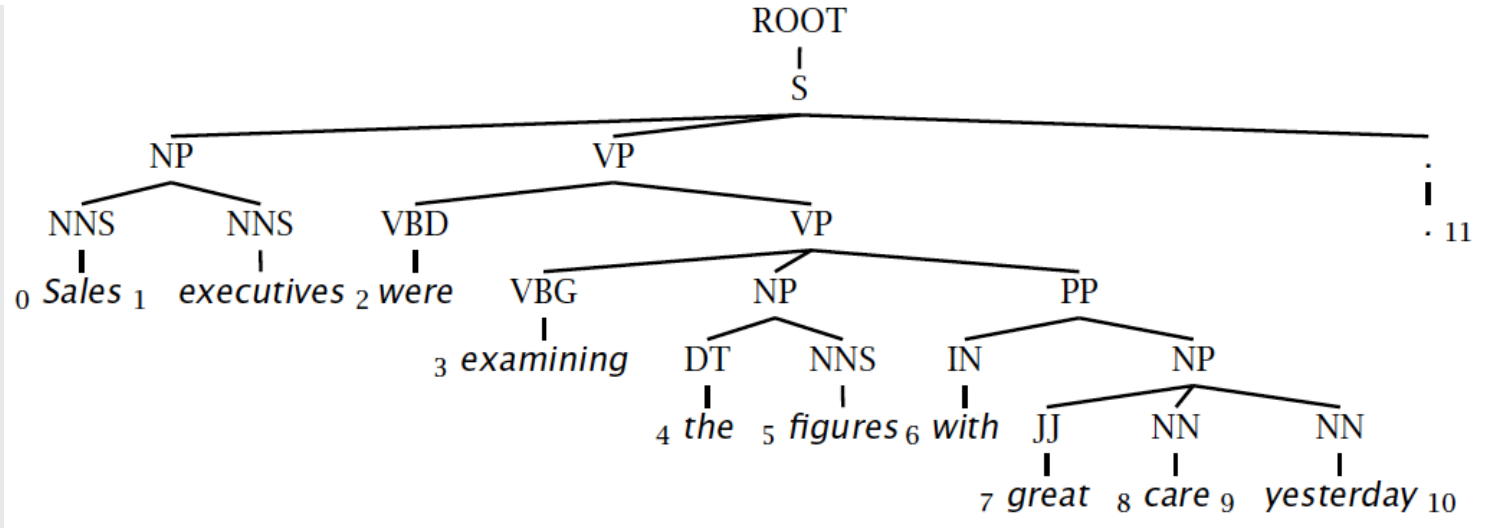
ОЦЕНКА

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)



[Chris Manning]

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%

[Chris Manning]

Прогресс

Parser	*F1 ≤ 40 words*	*F1 all words*
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**

[Chris Manning]

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

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: $\beta = \emptyset$

[Chris Manning]

Пример

1. Left-Arc$_r$ $\sigma|w_i, w_j|\beta, A \rightarrow \sigma, w_j|\beta, A \cup \{r(w_j, w_i)\}$
   Precondition: $(w_k, r', w_i) \notin A$, $w_i \neq$ 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: $(w_k, r', w_i) \in A$
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 …]	∅
Shift	[ROOT, я]	[пришел, к, …]	∅
LA$_{nsubj}$	[ROOT]	[пришел, к, …]	{nsubj (пришел, я)} = A_1
RA$_{root}$	[ROOT, пришел]	[к, тебе, …]	$A_1 \cup$ {root(ROOT, пришел)} = A_2
RA$_{prep}$	[ROOT, пришел, к]	[тебе, с, …]	$A_2 \cup$ {prep(пришел, к)} = A_3
RA$_{pobj}$	[ROOT, пришел, к, тебе]	[с, приветом, .]	$A_3 \cup$ {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, .]	$A_4 \cup$ {prep(пришел, с)} = A_5
RA$_{pobj}$	[ROOT, пришел, с, приветом]	[.]	$A_5 \cup$ {pobj(с, приветом)} = A_6
Reduce	[ROOT, пришел, с]	[.]	A_6
Reduce	[ROOT, пришел]	[.]	A_6
RA$_{punc}$	[ROOT, пришел, .]	[]	$A_6 \cup$ {(punc(пришел, .)} = A_7

по мотивам [Chris Manning]

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

[Chris Manning]

Оценка

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

UAS = 5 / 6 = 83%
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

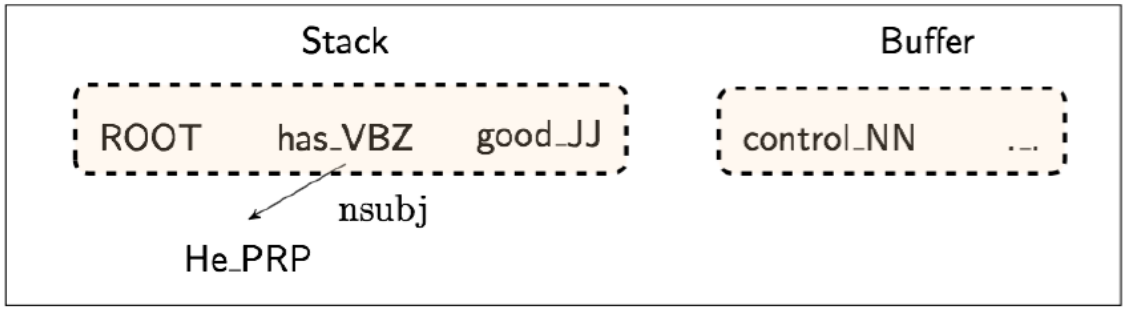
Разбор

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

по мотивам [Chris Manning]

Нейросетевой парсер

- We extract a set of tokens based on the stack / buffer positions:

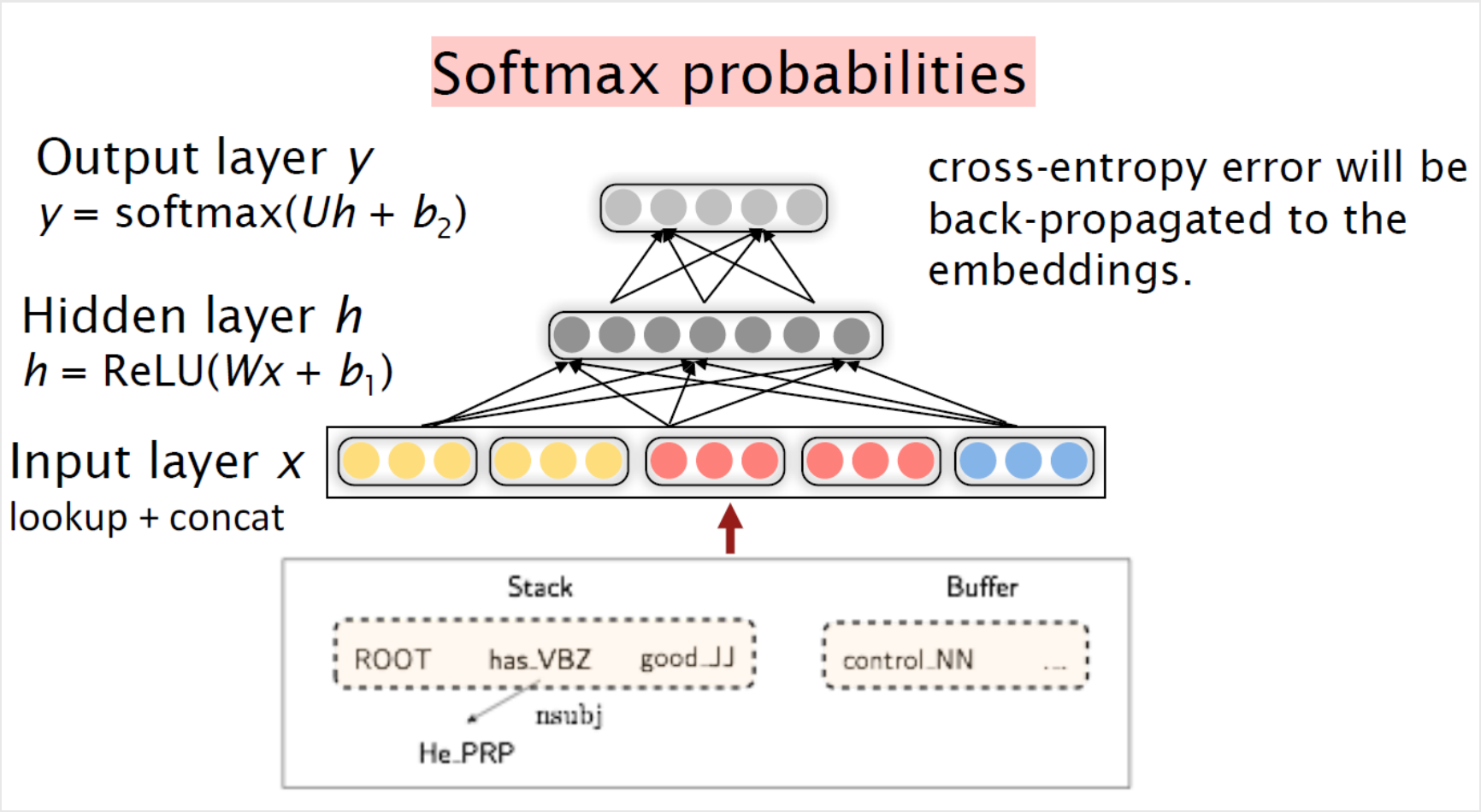


	word	POS	dep.
s_1	good	JJ	∅
s_2	has	VBZ	∅
b_1	control	NN	∅
$lc(s_1)$	∅	∅	∅
$rc(s_1)$	∅	∅	∅
$lc(s_2)$	He	PRP	nsubj
$rc(s_2)$	∅	∅	∅

- We convert them to vector embeddings and concatenate them

[Chen & Manning 2014] , слайды Manning & Socher

Архитектура



[Chen & Manning 2014] , слайды Manning & Socher

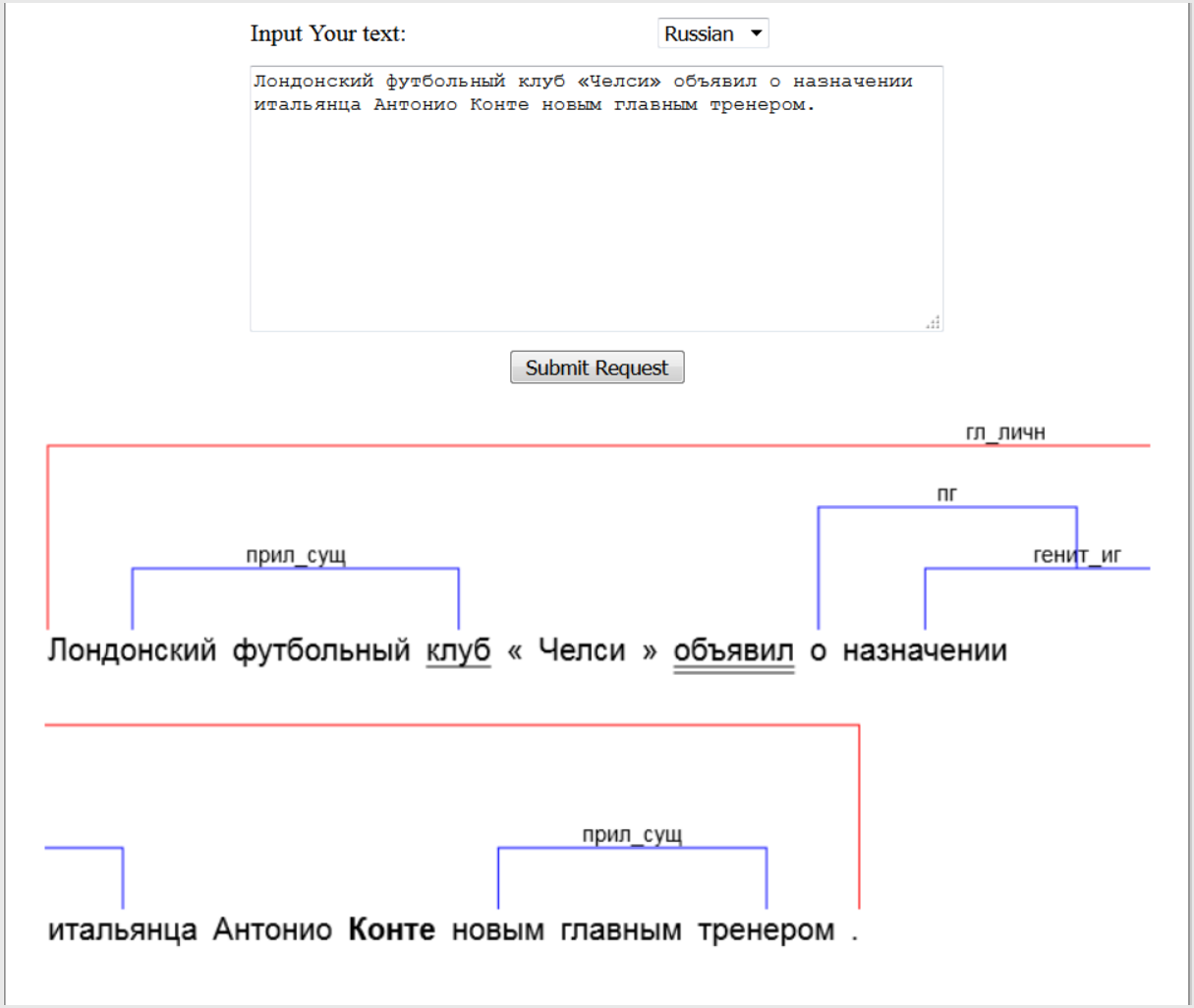
Оценка

Parser	UAS	LAS	sent. / s
MaltParser	89.8	87.2	469
MSTParser	91.4	88.1	10
TurboParser	**92.3***	89.6*	8
C & M 2014	92.0	**89.7**	**654**

1/30/18

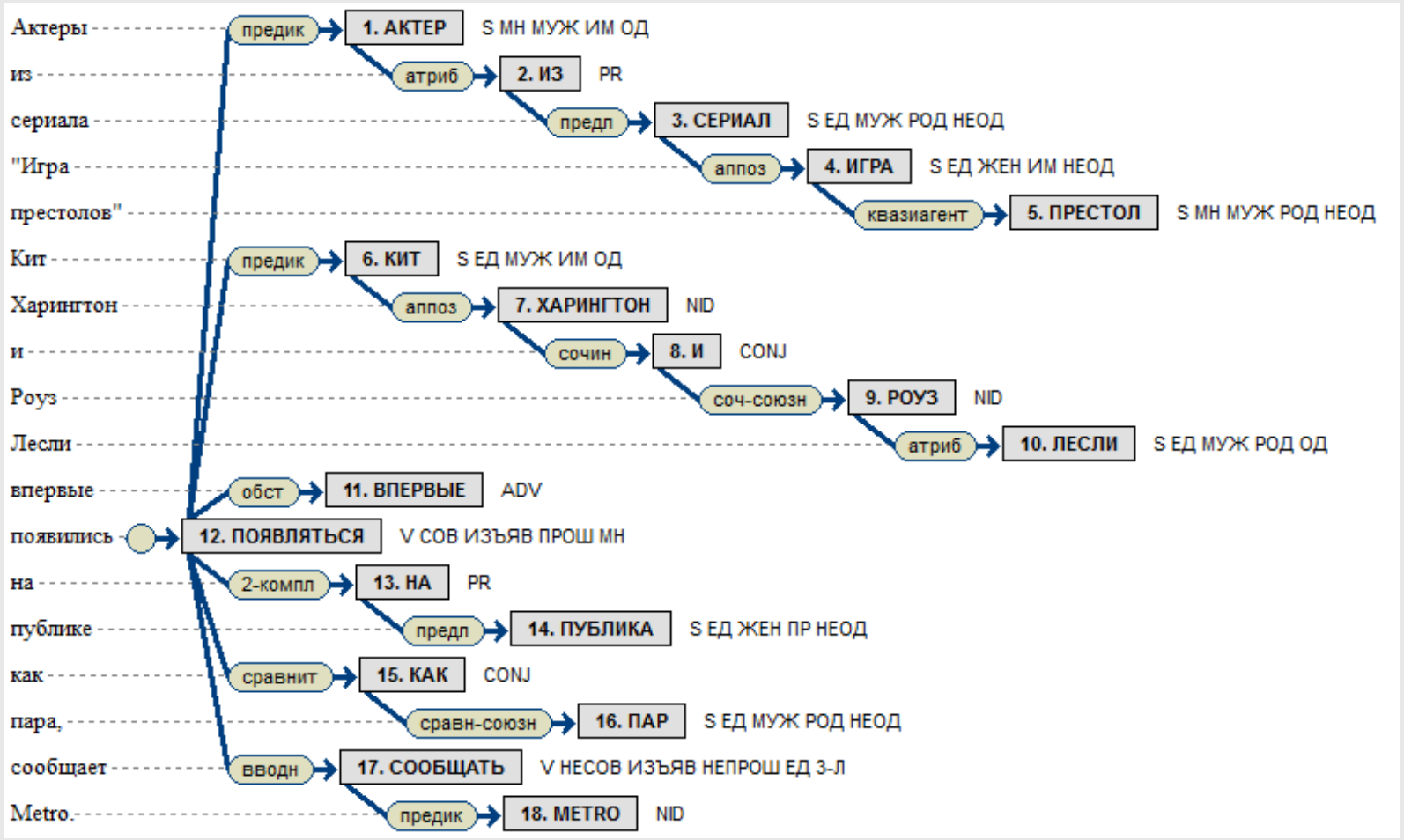
ПАРСЕРЫ ДЛЯ РУССКОГО

AOT



http://www.aot.ru/docs/synan.html

ЭТАП-3



http://proling.iitp.ru/ru/etap3

Томита-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)

https://tech.yandex.ru/tomita/

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

- feature embeddings
- beam search
- global normalization

Method	WSJ		Union-News		Union-Web		Union-QTB	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
Martins et al. (2013)*	92.89	90.55	93.10	91.13	88.23	85.04	94.21	91.54
Zhang and McDonald (2014)*	93.22	91.02	93.32	91.48	88.65	85.59	93.37	90.69
Weiss et al. (2015)	93.99	92.05	93.91	92.25	89.29	86.44	94.17	92.06
Alberti et al. (2015)	94.23	92.36	94.10	92.55	89.55	86.85	94.74	93.04
Our Local (B=1)	92.95	91.02	93.11	91.46	88.42	85.58	92.49	90.38
Our Local (B=32)	93.59	91.70	93.65	92.03	88.96	86.17	93.22	91.17
Our Global (B=32)	**94.61**	**92.79**	**94.44**	**92.93**	**90.17**	**87.54**	**95.40**	**93.64**
Parsey McParseface (B=8)	-	-	94.15	92.51	89.08	86.29	94.77	93.17

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%

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

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

http://www.dialog-21.ru/evaluation/2012/syntax/

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