

Морфологическая и синтаксическая разметка

Е.Ю. Калинина
МГУ, 2007-2008



Морфологическая разметка

Синонимы:

*part-of-speech tagging (POS-tagging),
частеречная разметка.*

Элементы данных морфологической
разметки включают:

- лемму;
- признак части речи;
- признаки грамматических категорий.



Морфологическая разметка: грамлеммы

(на основе системы ДИАЛИНГ) (1)

| <i>Attributes "pos" of the tag <ana></i> | <i>Attributes "gram" of the tag <ana></i> |
|--|---|
| <p>С - существительное, П - прилагательное, Г - глагол в личной форме, ПРИЧАСТИЕ - причастие ; ДЕЕПРИЧАСТИЕ – деепричастие, ИНФИНИТИВ – инфинитив, МС - местоимение- существительное, МС-П - местоименное прилагательное , МС-ПРЕДК - местоимение- предикатив , ЧИСЛ - числительное (количественное), ЧИСЛ-П - порядковое числительное,</p> | <p>мр, жр, ср - мужской, женский, средний род; од, но - одушевленность, неодушевленность; ед, мн - единственное, множественное число; им, рд, дт, вн, тв, пр, зв - падежи: именительный, родительный, дательный, винительный, творительный, предложный, звательный; 2 - второй родительный или второй предложный падежи; св, нс - совершенный, несовершенный вид; пе, нп - переходный, непереходный глагол; дст, стр - действительный, страдательный залог; нст, прш, буд - настоящее, прошедшее, будущее время;</p> |



Морфологическая разметка: грамлеммы (на основе системы ДИАЛИНГ) (2)

| <i>Attributes "pos" of the tag <ana></i> | <i>Attributes "gram" of the tag <ana></i> |
|---|--|
| <p>Н - наречие, ПРЕДК - предикатив, ПРЕДЛ - предлог, СОЮЗ - союз, МЕЖД - междометие, ЧАСТ - частица, ВВОДН - вводное слово, дфст - слово обычно не имеет множественного числа, опч - частая опечатка или ошибка, жарг, арх, проф - жаргонизм, архаизм, профессионализм, аббр – аббревиатура, безл - безличный глагол.</p> | <p>пвл - повелительная форма глагола; 1л, 2л, 3л - первое, второе, третье лицо; 0 - неизменяемое. кр - краткость (для прилагательных и причастий). сравн - сравнительная форма (для прилагательных). имя, фам, отч - имя, фамилия, отчество. лок, орг - локативность, организация. кач - качественное прилагательное. вопр, относ - вопросительность и относительность (для наречий).</p> |



Пример морфологической разметки (на основе системы ДИАЛИНГ)

```
<?xml version="1.0" encoding="windows-1251" ?> <text> <p>  
<s>  
<w>Звонили<ana lemma="ЗВОНИТЬ" pos="Г" gram="мн,нс,нп,дст,  
прш," /></w>  
<w>к<ana lemma="К" pos="ПРЕДЛ" gram="" /></w>  
<w>вечерне  
<ana lemma="ВЕЧЕРНЯ" pos="С" gram="жр,ед,дт,пр,но," />  
<ana lemma="ВЕЧЕРНИЙ" pos="П" gram="ср,ед,кр," /></w>  
<pun>.</pun> </s>  
<s><w>Торжественный<ana lemma="ТОРЖЕСТВЕННЫЙ" pos="П"  
gram="мр,ед,им,вн," /></w>  
<w>гул<ana lemma="ГУЛ" pos="С" gram="мр,ед,им,вн,но,"  
></w>  
<w>колоколов  
<ana lemma="КОЛОКОЛ" pos="С" gram="мр,мн,рд,но," />  
<ana lemma="КОЛОКОЛОВ" pos="С" gram="мр,фам,ед,им,од,"  
></w>  
.....<pun>.</pun> </s></p></text>
```



Принципы разметки

- Описание (обоснование) схемы разметки
- Общепринятая система лингвистических понятий
- Известная для пользователя схема анализа
- Мотивированность введения параметров
- Теоретически нейтральная (традиционная) схема разметки



Проблемы морфологического анализа и морфологической разметки: омонимия (1)

- {s}Я{|я=S,ед,од=им,жен|я=S,ед,од=им,муж} **сидел**{сидеть=V,несов=прош,ед,изъяв,муж} **на**{на=PART=|на=PR=}
- **барском**{барский=A=пр,ед,муж|барский=A=пр,ед,сред}
сиденье{сиденье=S,сред,неод=им,ед|сиденье=S,сред,неод=вин,ед|сиденье=S,сред,неод=пр,ед}, **дышал**{дышать=V,несов=прош,ед,изъяв,муж} **горячим**{горячий=A=дат,мн|горячий=A=твор,ед,муж|горячий=A=твор,ед,сред|горячее=S,ед,сред,неод=твор|горячить=V,несов=непрош,ед,прич,кр,муж,страд|горячить=V,несов=непрош,мн,изъяв,1-л} **ветром**{ветер=S,муж,неод=твор,ед}, **бившим**{бить=V,несов=прош,дат,мн,прич|бить=V,несов=прош,твор,ед,прич,муж|бить=V,несов=прош,твор,ед,прич,сред} **в**{в=PR=} **лицо**{лицо=S,сред,неод=им,ед|лицо=S,сред,неод=вин,ед|лицо=S,сред,од=им,ед|лицо=S,сред,од=вин,ед}, **ощущая**{ощущать=V=непрош,деепр,несов} **в**{в=PR=}



Проблемы морфологического анализа и морфологической разметки: омонимия (2)

- **то**{то=CONJ=|тот=A=им,ед,сред|тот=A=вин,ед,сред|то=S,ед,сред,неод=им|то=S,ед,сред,неод=вин} **же**{же=PART=|же=CONJ=} **время**{время=S,сред,неод=им,ед|время=S,сред,неод=вин,ед} **не**{не=PART=} **истребимую**{истребимый=A=вин,ед,жен} **никакими**{никакой=A=твор,мн} **сквозняками**{сквозняк=S,муж,неод=твор,мн} **пыль**{пыль=S,ед,жен,неод=им|пыль=S,ед,жен,неод=вин} **и**{и=PART=|и=INTJ=|и=CONJ=} **легкий**{легкий=A=им,ед,муж|легкий=A=вин,ед,муж,неод}
- **запах**{запах=S,муж,неод=им,ед|запах=S,муж,неод=вин,ед|запах=S,муж,неод=им,ед|запах=S,муж,неод=вин,ед|запахнуть=V,сов=прош,ед,изъяв,муж} **духов**{духов=A=им,ед,муж|духов=A=вин,ед,муж,неод|дух=S,муж,неод=род,мн|дух=S,муж,од=род,мн|дух=S,муж,од=вин,мн|духи=S,мн,муж,неод=род} --



Проблемы морфологического анализа и морфологической разметки

- незнакомые слова: Махабхарата, фотосправочник, короткохоботый
- экзотические формы: лузях (С.Соколов), вспорхливый, творческ, почил в бозе,
- авторские варианты написания: итти, казалось, бодростию



Сложные лексические единицы: чему приписываем тэг?

- Наречия: без удержу, до отвала, с гаком
- Вводные слова: в сущности, между прочим
- Союзы: коль скоро, лишь бы, даром что
- Предлоги: в преддверии, вплоть до
- Частицы: все ж таки, как бы не так
- Фразеологические предикативы (?): кот наплакал; раз, два и обчелся etc.



Морфологический стандарт Русского национального корпуса

- Лексема, которой принадлежит словоформа (указывается «словарная запись» данной лексемы и ее принадлежность к той или иной части речи).
- Множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (например, род для существительного, переходность для глагола).
- Множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола).
- Информация о нестандартности грамматической формы, орфографических искажениях и т. п.



Морфологический стандарт русского национального корпуса: исходная лексема

- Для всех словоформ, принадлежащих видовым парам, указываются исходные формы обоих видов (например, форма *пришёл* считается принадлежащей и лексеме ПРИЙТИ, и лексеме ПРИХОДИТЬ).
- Для -ся-форм в тех случаях, когда существуют лексемы с -ся и без -ся, указываются обе исходные формы (например, форма *разрушается* считается принадлежащей и лексеме РАЗРУШАТЬСЯ, и лексеме РАЗРУШАТЬ).
- Для прилагательных, совпадающих с причастиями (*открытый*), в качестве исходной дается как лексема-прилагательное (ОТКРЫТЫЙ), так и глагол (ОТКРЫТЬ).



Морфологический стандарт русского национального корпуса: части речи

- **S** — существительное (*яблоня, лошадь, корпус, вечность*)
S-PRO — местоимение-существительное (*она, что*)
A — прилагательное (*коричневый, таинственный, морской*)
A-PRO — местоимение-прилагательное (*который, твой*)
NUM — числительное (*четыре, десять, много*)
A-NUM — числительное-прилагательное (*один, седьмой, восьмидесятый*)
PRAEDIC — предикатив (*жаль, хорошо, пора*)
A-PRAEDIC — местоимение-предикатив (*некого, нечего*)
V — глагол (*пользоваться, обрабатывать*)
ADV — наречие (*сгоряча, очень*)
ADV-PRO — местоименное наречие (*где, вот*)
PR — предлог (*под, напротив*)
CONJ — союз (*и, чтобы*)
PART — частица (*бы, же, пусть*)
INTJ — междометие (*увы, батюшки*)
PARENTH — вводное слово (*кстати, по-моему*)



Морфологический стандарт русского национального корпуса: грамматические категории

- **Падеж:**
- **nom** — именительный падеж (*голова, сын, степь, сани, который*)
- **gen** — родительный падеж (*головой, сына, степи, саней, которого*)
- **acc** — винительный падеж (*голову, сына, степь, сани, который/которого*)
- **dat** — дательный падеж (*голове, сыну, степи, саням, которому*)
- **loc** — предложный падеж (*[о] голове, сыне, степи, санях, котором*)
- **ins** — творительный падеж (*головой, сыном, степью, санями, которым*)
- **gen2** — второй родительный падеж (*чашка чаю*)
- **acc2** — второй винительный падеж (*постричься в монахи; по два человека*)
- **loc2** — второй предложный падеж (*в лесу, на оси*)
- **voc** — звательная форма (*Господи, Серёж, ребят*)



Морфологический стандарт русского национального корпуса: грамматические категории

- **Степень сравнения:**
- **comp** — сравнительная степень
(*глубже*)
comp2 — форма
«по+сравнительная степень»
(*поглубже*)
supr — превосходная степень
(*глубочайший*)



Морфологический стандарт русского национального корпуса: грамматические категории

- **Залог:**
- **act** — действительный залог (*разрушил, разрушивший*)
- **pass** — страдательный залог (только у причастий: *разрушаемый, разрушенный*)
- **med** — медиальный, или средний залог (глагольные формы на *-ся*: *разрушился* и т.п.)



Морфологический стандарт русского национального корпуса: другие множественные пометы

- В ряде случаев допускается множественная помета части речи для союзов/частиц типа *словно*, для *-о/-е*-форм типа *хорошо* (предикатив/наречие/прилагательное), для субстантивированных адъективов типа *всё*, *военный* (существительное/прилагательное), для форм *его*, *её*, *их* (притяжательное/личное местоимение); число таких случаев по мере работы над корпусом будет уменьшаться.
- Ставится множественная помета в случаях, когда выбор лексемы или грамматического значения невозможен (*не видел родного отца* — **gen/acc**; *манекену* — **anim/inan**; *спазмами* — исходная форма СПАЗМ/СПАЗМА, и т. п.)



Морфологический стандарт русского национального корпуса: дополнительные пометы

- **anom** («Аномальная форма») — различного рода морфологические аномалии, возможные у устаревших или просторечных нелитературных форм (*три дни, ляжь*)
- **distort** («Искаженная форма») — орфографическое и/или фонетическое искажение слова, передающее различные особенности произношения (*дэвушка, това'ищи, про-хо-ди, низнаю*), а также сокращения (не аббревиатуры) и иные особенности записи (*тов., 1-й*).
- Кроме того, в корпусе с неснятой омонимией используется особая помета (**bastard**) для формы, порожденной автоматическим анализатором по аналогии: например, форма вроде *Махабхарата* получает несколько гипотетических разборов, в том числе от псевдолексем *махабхаронок, махабхарать* и т.



Морфологическая разметка BNC

- *could've* = <w VM0>could<w VHI>'ve
doesn't = <w VDZ>does<w XX0>n't
dunno = <w VDB>du<w XX0>n<w VVI>no
wanna = <w VVB>wan<w TO0>na
--or-- <w VVB>wan<w AT0>na
gimme = <w VVB>gim<w PNP>me



Морфологическая разметка BNC

- *<w AV0>of course* (adverb)
- *<w PRP>according to* (preposition)
- *<w NN1>persona non grata*
('naturalised' compound noun)
- *<w CJS>except that* (conjunction)



Морфологическая разметка BNC

- *she <w VBZ>is playing her best tennis for six years.*
[CH3.1383]
- *she <w VBZ>is just a star.*
[CH3.6940]
- *John <w VHZ>has built a set of bookshelves.*
[C9X.121]
- *John <w VHZ>has great courage.*
[CA9.1941]
- *We <w VDD>did<w XX0>n't see anybody.*
[KB2.702]
- *They <w VDB>do nice work.*



Морфологическая разметка BNC

- *We* <w VM0>*can go there.*
- *We* <w VM0>*could go there.*
- *We* <w VM0>*used* <w TO0>*to go there every year.*
- The form *let's* is treated as one verb:
- <w VM0>*Let's* <w VVI>*go!*



Морфологическая разметка BNC

- **Subjunctives** and **Imperatives**. (Both take V-B tags)
- *She suggested that they <w VVB>get married.*
[CBC.12107]
- *Please <w VBB>be patient.*
[CHJ.901]
- *<w VDB>Do<w XX0>n't just stand there watching!*
[ACB.3470]



Морфологическая разметка BNC

- **Catenative** or **semi-auxiliary** verbs such as *going to*, *ought to*, and *used to* + infinitive
- *we're* <w VVG>*going* <w TO0>*to get killed*.
[HNN.445]
- *you* <w VM0>*ought* <w TO0>*to let them know*.
[KCT.6117]



Морфологическая разметка BNC

- **ADJECTIVE vs. ADVERB**
- *We arrived <w AJ0>tired, but <w AJ0>safe*
[CCP.530]
- *Peter sang out <w AV0>loud and <w AV0>clear.*



Морфологическая разметка BNC

- **ADJECTIVE vs. NOUN**
- *a <w AJ0>white screen, The screen is <w AJ0>white.*
- *<w NN1>Red is my favourite colour.*
- *They painted the wall a brilliant <w NN1>white.*
- *two <w AJ0>smiling children ('two children who are/were smiling')*
[HTT.743]
- *new <w NN1>spending plans ('new plans for spending')*
- *his <w NN1>reading ability ('his ability in reading')*



Морфологическая разметка BNC

- **ADJECTIVE vs. VERB**
- *The effect is <w AJ0>lasting (compare a <w AJ0>lasting effect).*
- *The door is <w AJ0>locked (compare the <w AJ0>locked door.)*
- *The man was <w VVG>dying.
[HTM.1494 *VVG-AJ0]*
- *BUT: the <w AJ0>dying man.
[FSH.606]*
- *an <w NN1>interest <w VVG>earning account*
- *a <w NN1>hypothesis <w VVN>driven approach*



Синтаксическая разметка

- фиксация синтаксических связей
- приписывание синтаксическим единицам соответствующих характеристик:
 - тип предложения
 - синтаксическая функция
 - член предложения
 - и т.п.



Сложности синтаксической разметки: разнообразии синтаксических теорий и формализмов:

- грамматика зависимостей;
- грамматика непосредственно-составляющих;
- грамматика структурных схем;
- традиционные синтаксические учения о членах предложения;
- грамматика конструкций;
- лексико-функциональная грамматика (LFG) и др.

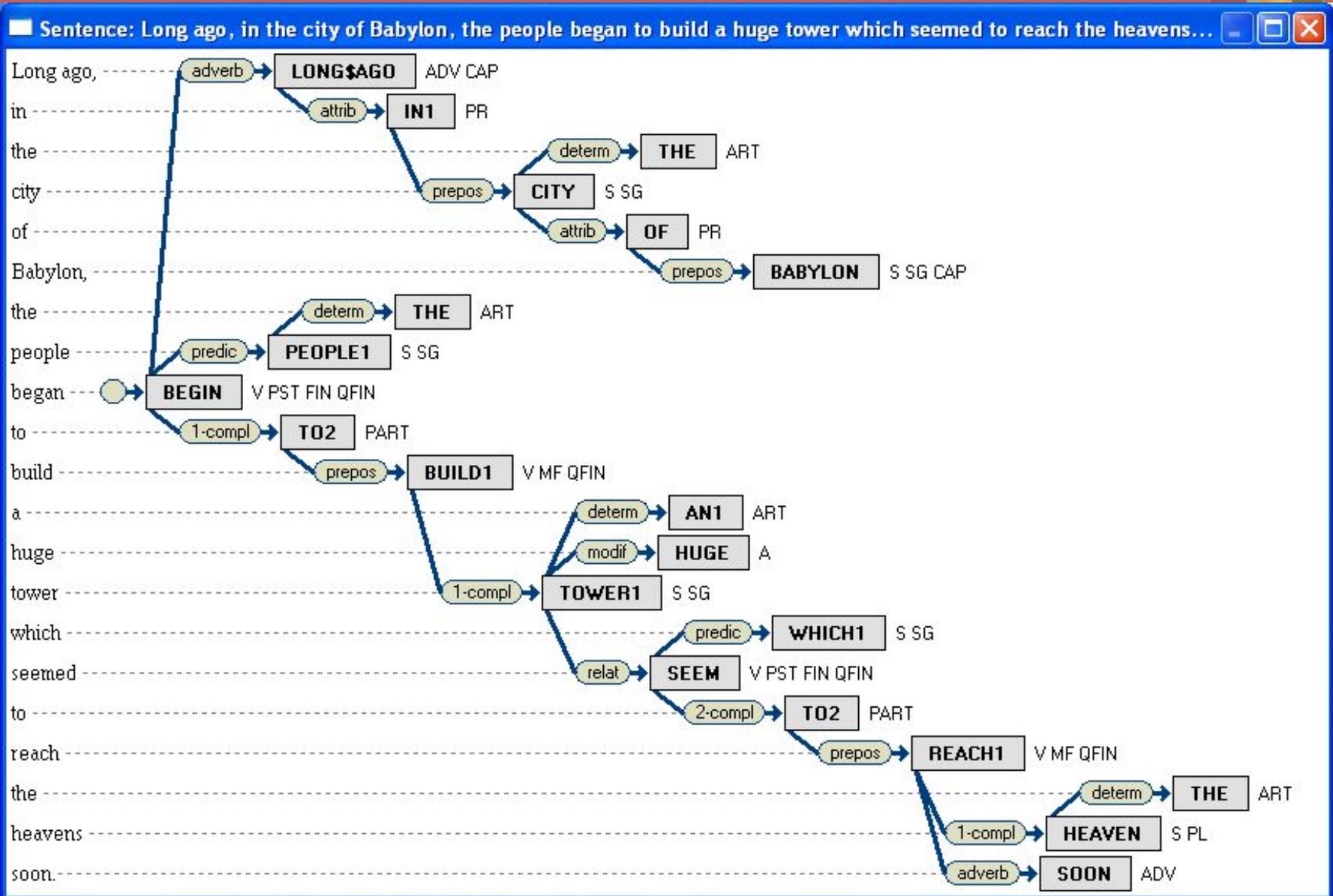


Пример синтаксического разбора (грамматика зависимостей, система ЭТАП-3)

*Long ago, in the city of Babylon, the
people began to build a huge tower
which seemed to reach the heavens
soon.*



Пример синтаксического разбора



Penn Tree Bank

- **The Penn Treebank syntactic tagset**
- 1. ADJP Adjective phrase
- 2. ADVP Adverb phrase
- 3. NP Noun phrase
- 4. PP Prepositional phrase
- 5. S Simple declarative clause
- 6. SBAR Clause introduced by subordinating conjunction or 0 (see below)
- 7. SBARQ Direct question introduced by wh-word or wh-phrase
- 8. SINV Declarative sentence with subject-aux inversion
- 9. SQ Subconstituent of SBARQ excluding wh-word or wh-phrase
- 10. VP Verb phrase
- 11. WHADVP Wh-adverb phrase
- 12. WHNP Wh-noun phrase
- 13. WHPP Wh-prepositional phrase
- 14. X Constituent of unknown or uncertain category



Penn Tree Bank

- Null elements
- 1. * ``Understood" subject of infinitive or imperative
- 2. 0 Zero variant of that in subordinate clauses
- 3. T Trace---marks position where moved wh-constituent is interpreted
- 4. NIL Marks position where preposition is interpreted in pied-piping contexts

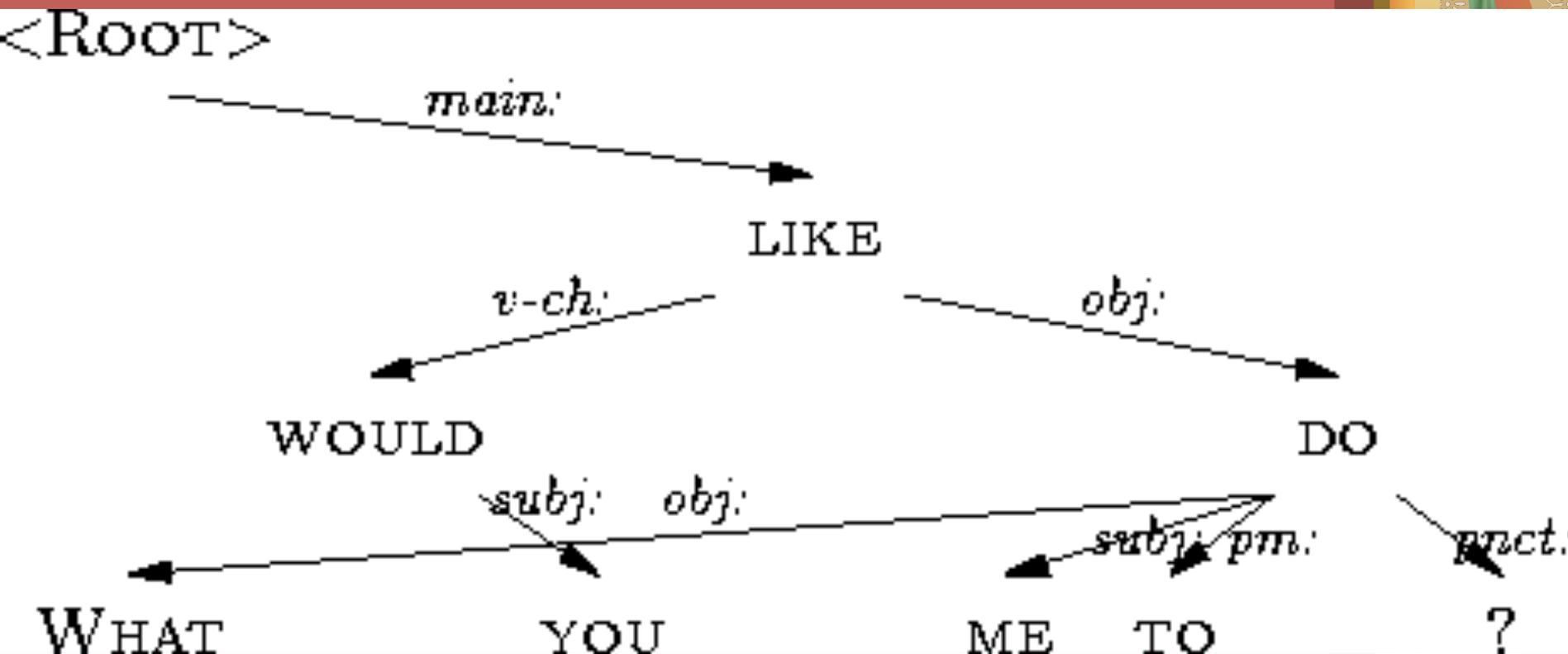


Penn Tree Bank

- Functional tags
- CLF – true clefts
- NOM – non NPs that function as NPs
- ADV – clausal and NP adverbials
- LGS – logical subjects in passive constructions
- PRD – non-VP predicates
- SBJ – logical subjects
- TPC – topicalized and fronted constituents



Дерево зависимостей: Connexor



Дерево зависимостей: Connexor

```
<What>  
  "what" <**CLB> PRON WH SG/PL @OBJ  
<would>  
  "would" V AUXMOD VFIN @+FAUXV  
<you>  
  "you" PRON PERS NOM SG2/PL2 @SUBJ  
<like>  
  "like" V INF @-FMAINV  
<me>  
  "i" PRON PERS ACC SG1 @OBJ  
<to>  
  "to" INFMARK> @INFMARK>  
<do>  
  "do" V INF @-FMAINV  
<?>
```



Семантическая разметка

- Аргентина **идет** русским путем ...
- Игорь Трунов тут же пояснил, что речь **идет** об одном миллионе долларов. ...
- Неужели Соколов не понимает, что речь **идет** о чем-то неизмеримо большем, чем о ...
- Кредитование реального сектора **идет** ни шатко ни валко. ...
- Как подтвердил "Известиям" Эдуард Кузьмин, все **идет** по плану ...
- Россия -- страна, которая **идет** к открытому обществу и не боится ...
- ...что, во-первых, о моей режиссуре и речи не **идет**, и, во-вторых, как актер я ...
- Судьба ведет человека, но человек **идет** потому, что хочет, и он волен не хотеть...
- И вот уже ребенок **идет** от лужи, идет с чужим дядей, ...
- ... звенело в ушах и все казалось, эшелон **идет**, идет.. ...
- ... он, убитый, все жал на акселератор, и танк **идет**.
- Впрочем, речь **идет** не обо мне...



Таксономическая разметка НКРЯ

- *Ничего*{ничто=М-С, ср, ед=рд} *общего*
{общий=П=ср, ед, рд, **Class="соц_отн-я"**
| **Class="охват"**]} *с*{с=ПРЕД}
европейскими{европейский=П=мн, тв}
акварелями{акварель=С, жр, но=мн, тв
Class="в-во" | **Class="изображение"**]}
Наматжиры{Наматжира*=С, фам, мр,
од=ед, рд} *и*{и=СОЮЗ} *его*{его=М-П}
последователей{последователь=С, мр,
од=мн, вн **Class="человек"**]} (Даниил
Гранин. Месяц вверх ногами)



Таксономическая разметка НКРЯ

кузов

- класс = емкость
- мереологический класс = часть
- мереологический коррелят = транспортное средство
- семантическая одушевленность = неодушевленное

интриганка

- класс = лицо
- пол = женский
- оценка = отрицательная
- семантическая одушевленность = одушевленное
- деривационный класс = nomina feminina



Таксономическая разметка, GNOME

This table's

```
<ne id="ne2" cat="poss-np" per="per3" num="sing"
```

- gen="neut" gf="subj" lftype="term"
- onto="concrete" ani="inanimate"
- deix="deix-no" count="undersp-count"
- generic="generic-no">

```
<ne id="ne3" cat="this-np" per="per3" num="sing"
```

- gen="neut" gf="gen" lftype="term"
- onto="concrete" ani="inanimate"
- deix="deix-yes" count="count-yes"
- structure="atom"
- generic="generic-no">



Таксономическая разметка, GNOME

(allow)

- `<ne id="ne4" cat="bare-np" per="per3" num="plur"`
- `gen="neut" gf="obj" lftype="term" onto="person"`
- `ani="animate" deix="deix-no" count="count-yes"`
- `structure="set" generic="generic-yes">`

scholars `</ne>`

(to link)

- `<ne id="ne5" cat="pers-pro" per="per3"`
`num="sing"`
- `gen="neut" gf="obj" lftype="term" onto="concrete"`
- `ani="inanimate" deix="deix-yes" count="count-yes"`
- `structure="atom" generic="generic-no"> it </ne>`



Семантическая разметка: ОНТОЛОГИИ

- And 00000000 the 00000000
- soldiers 23241000platted 21072000
- a 00000000crown 21110400
- of 00000000thorns 13010000
- and 00000000put 21072000
- it 00000000on 00000000
- his 00000000head 21030000
- and 00000000they 00000000
- put 21072000on 00000000
- him 00000000a 00000000
- purple 31241100robe 21110321



Семантическая разметка:

ОНТОЛОГИИ

- 00000000 Low content word (and, the, a, of, on, his, they etc)
- 13010000 Plant life in general
- 21030000 Body and body parts
- 21072000 Object-oriented physical activity (e.g. put)
- 21110321 Men's clothing: outer clothing
- 21110400 Headgear
- 23231000 War and conflict: general
- 31241100 Colour



Семантическая разметка Penn Tree Bank

- *Vandenberg and Rayburn were wise enough *TRACE* to leave specific operations to presidents.*
- base=leave2; tense=infinitival;
- arg2=presidents;
- arg1=specific operations;
- arg0=**TRACE** -> Vandenberg and Rayburn;



Семантическая разметка Penn Tree Bank

- **HIT** (sense: strike)
- Arg0: hitter
- Arg1: thing hit
- Arg2: instrument, hit with
- **HAIL** (sense: pellets of ice from the sky)
- Labels allow to capture transitivity alternations:
- *John (Arg0) broke the window*
- *(Arg1) and The window (Arg1) broke.*



Семантическая разметка Penn Tree Bank

- **EDGE** (sense: move slightly)
- Arg0: causer of motion³
- Arg1: thing in motion
- Arg2: distance moved
- Arg3: start point
- Arg4: end point
- Arg5: direction
- *The publishing unit reported revenue edged up 2.6% to \$263.2 million from \$256.6 million.*



Семантическая разметка Penn Tree Bank

- **BUY**
- Arg0: buyer
- Arg1: thing bought
- Arg2: seller, bought-from
- Arg3: price paid
- Arg4: benefactive, bought-for



Семантическая разметка Penn Tree Bank

- **PURCHASE** **BUY** **SELL**
- Arg0: buyer Arg0: buyer Arg0: seller
- Arg1: thing bought Arg1: thing bought Arg1: thing sold
- Arg2: seller Arg2: seller Arg2: buyer
- Arg3: price paid Arg3: price paid Arg3: price paid
- Arg4: benefactive Arg4: benefactive Arg4: benefactive



Семантическая разметка Penn Tree Bank

The company bought a wheel-loader from Dresser.

- Arg0: The company
- rel: bought
- Arg1: a wheel-loader
- Arg2-from: Dresser

TV stations bought "Cosby" reruns for record prices.

- Arg0: TV stations
- rel: bought
- Arg1: "Cosby" reruns
- Arg3-for: record prices.



Семантическая разметка Penn Tree Bank

- LOC: location NEG: negation marker
- TMP: time MOD: modal verb
- MNR: manner EXT: extent, numerical role
- DIR: direction PRP: purpose
- CAU: cause ADV: general-purpose modifier



