



Синтаксически аннотированный корпус чешского языка

Аня Недолужко,
Ян Гаич и кол.



Синтаксически аннотированный корпус чешского языка

Prague Dependency Treebank (PDT)

ТЕКСТЫ НА ЧЕШСКОМ ЯЗЫКЕ

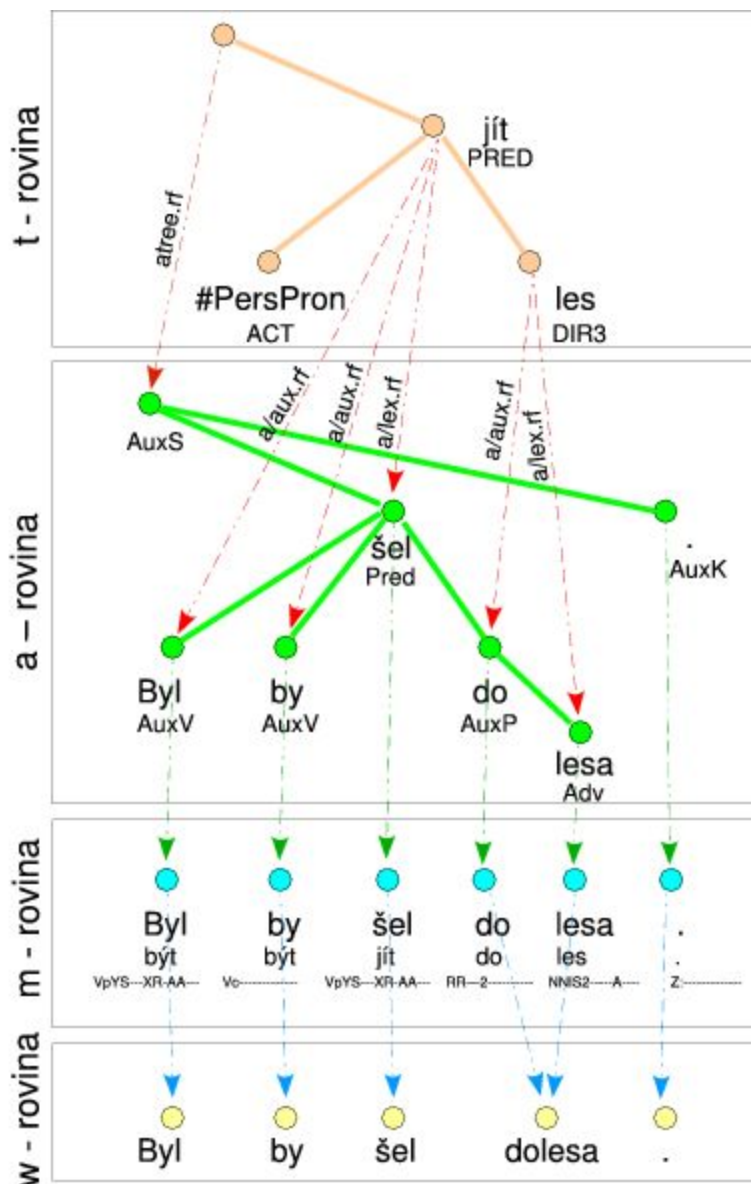
- морфологический уровень: 2 млн. слов
- поверхностно-синтаксический уровень: 1.5 млн. слов
- глубинно-синтаксический уровень: 0.8 млн. слов



***Byl by
šel
dolesa.***



***(Он)
пошёл
бы в лес.***





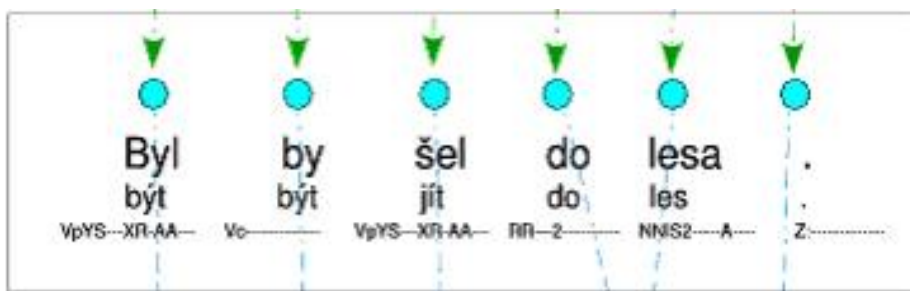
Морфологический уровень



(Он) шёл бы в лес.

атрибуты:

- атрибут lemma
- атрибут tag – 15 позиций,
напр. NNIS2-----A-----
- и др.

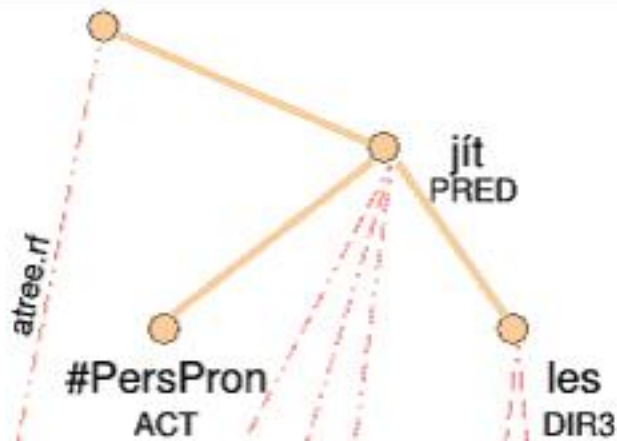




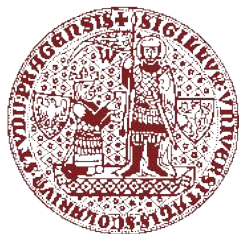
Глубинно- синтаксический уровень

атрибуты (39)

(Он) шёл бы в лес.



- functor (ACT, PAT, ADDR, PRED, DENOM, PAR, CONJ, LOC, DIR1, DIR2, TWHEN, TTILL и др.)
- t_lemma
- gram/sempos, gram/verbmod



Словарь моделей управления VALLEX



rozumět *impf* v

1 (*vyznat se; chápat*) ACT(1) PAT(3|zda|že|cont) ◇*rozumí úloze; nerozuměl, zda to má nebo nemá udělat; rozumíš už, co se stalo?; rozumí dobře anglicky; matka dceři rozumí* ✕ rcp: ACT-PAT; class: mental action

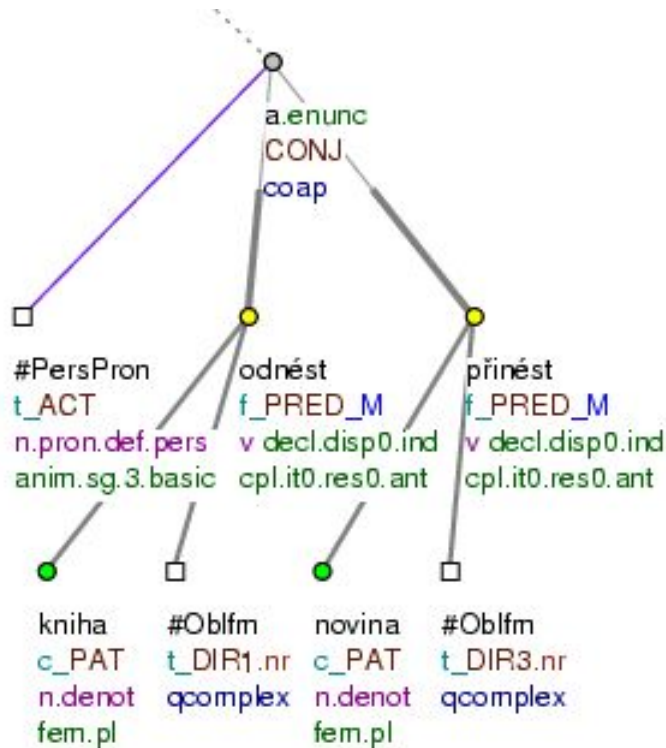
2 (*rozlišovat; znát*) ACT(1) ADDR(3) PAT(4|zda|že|cont) ěMANN ◇*rozumí mu každé slovo; Rozumím vám správně, že rozpočet v parlamentě podpoříte? (ČNK)* ✕ rcp: ACT-ADDR; class: communication

3 idiom (*chápat*) ACT(1) PAT(4|že) EFF(7|pod+7) ◇*rozumí tím / pod tím příměří* ✕ rfl: pass; class: mental action

cz. rozumět = ru. понимать



Актуальное членение



атрибуты:

- tfa (t, c, f)
- deepord (глубинный порядок узлов, основанный на функциональной перспективе предложения)

Kniha odnesl a noviny přinesl. - (Он) книги унес, а журналы принес.



Кореференция



- **грамматическая** (`coref_gram.rf`)
кореференция возвратных и относительных местоимений
- **текстовая** (`coref_text.rf`)
PersPron и PossPron 3-го лица,
DemPron *этот* в субст.функции, эллипсис
- **особые случаи** (`coref_special – exoph, segm`)



Расширенная аннотация кореференции (текстовая кореференция)

- **0** (повтор NP антецедента, пары Pron – NP, напр. *дом – (этот) дом, он – Петя* и др.)
- **SYN** (повтор – синоним антецедента, напр. *предприятие - завод*)
- **ER** (повтор – гипероним антецедента, напр. *Петя – этот ребенок*)
- **NR** (анафорическое отношение нереферентных NP)



Расширенная аннотация кореференции ("bridging anaphora")



- **PART** (отношение часть—целое, напр. *дверь* - *ручка*)
- **SET** (отношение множество-подмножество, элемент множества, напр. *мушкетёры* - *Атос*)
- **FUNCT** (отношение функциональной принадлежности, напр. *школа* - *учитель*)
- **CONTRAST** (отношение семантического противопоставления, напр. *взрослый* - *ребенок*)
- **REST** - другое



Другие проекты



- Prague Arabic Dependency Treebank,
<http://www ldc.upenn.edu>
- Prague Czech-English Dependency Treebank,
<http://ufal.mff.cuni.cz/pcedt>



Планируется...



- аннотация разговорных текстов,
- детализация имеющейся аннотации,
- аннотация типологически отличных языков,
- аннотация параллельных чешских и английских текстов
- разработка нового уровня аннотации



<http://ufal.mff.cuni.cz/>