

PDT 2.0

teorie a praxe

Vašek Němčík
(`xnemcik@fi.muni.cz`)

CZPJ, FI MU Brno

3. prosince 2010

Přehled

- Motivace
- PDT 2.0 v kostce
- Vybrané vlastnosti
- Možnosti použití PDT 2.0

Co je to PDT2?

- Prague Dependency Treebank
- korpus českých textů
- patrně největší lingvistický projekt v ČR
- bohatá ruční anotace, více rovin
- Institut formální a aplikované lingvistiky, Univerzita Karlova, Praha
- dostupné přes LDC

Důvody vzniku PDT2

- FDP (Funkční generativní popis) navržený Pražskou školou ...
ověření teorie v praxi, “na reálných datech”
- získání dat pro trénování metod strojového učení, pro NLP aplikace

PDT2 obsahuje ...

- 3 úrovně anotace
(zhruba) odpovídá FGP
 - **w-rovina** *slova*
 - **m-rovina** *morf. značky* 2 mil.
 - **a-rovina** *syntax* 1.5 mil.
 - **t-rovina** *hloubková syntax* 0.8 mil.
- každá rovina je navázána na nižší rovinu
- nikoliv 1:1, uzly přidávány/mazány

Důvody k použití PDT2?

pro češtinu unikátní s ohledem na:

- rozsah a bohatost anotace
- smysl pro detail
 - celá řada syntaktických jevů
 - valence
 - AČV (information structure; topic-focus)
 - koreference

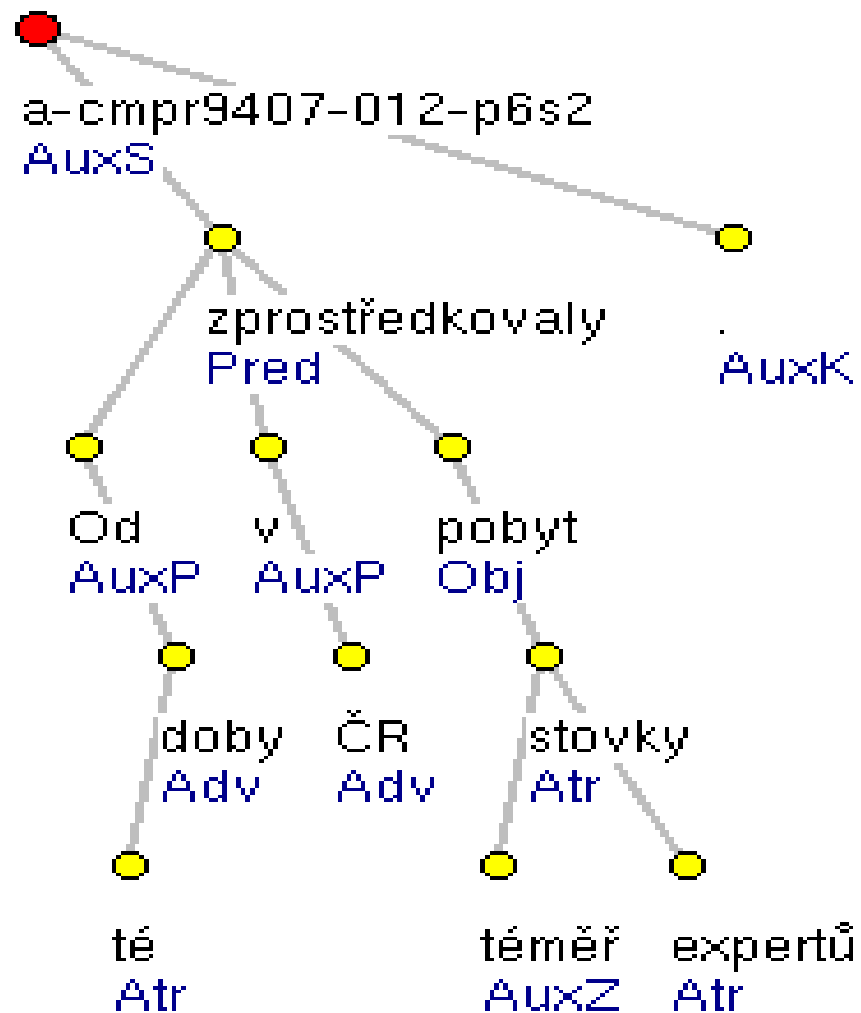
Proč PDT2 zajímá mne?

- anotace koreference (manuální)
 - v žádném jiném českém korpusu
 - rozsah cca. 45.000 koref. párů
- co potřebuji:
 - základní strukturu: věty, klause, slova
 - NP a zájmena (i nulová) + morf. kategorie
 - anaforické odkazy samotné
 - (v podstatě *jednoduché věci*)

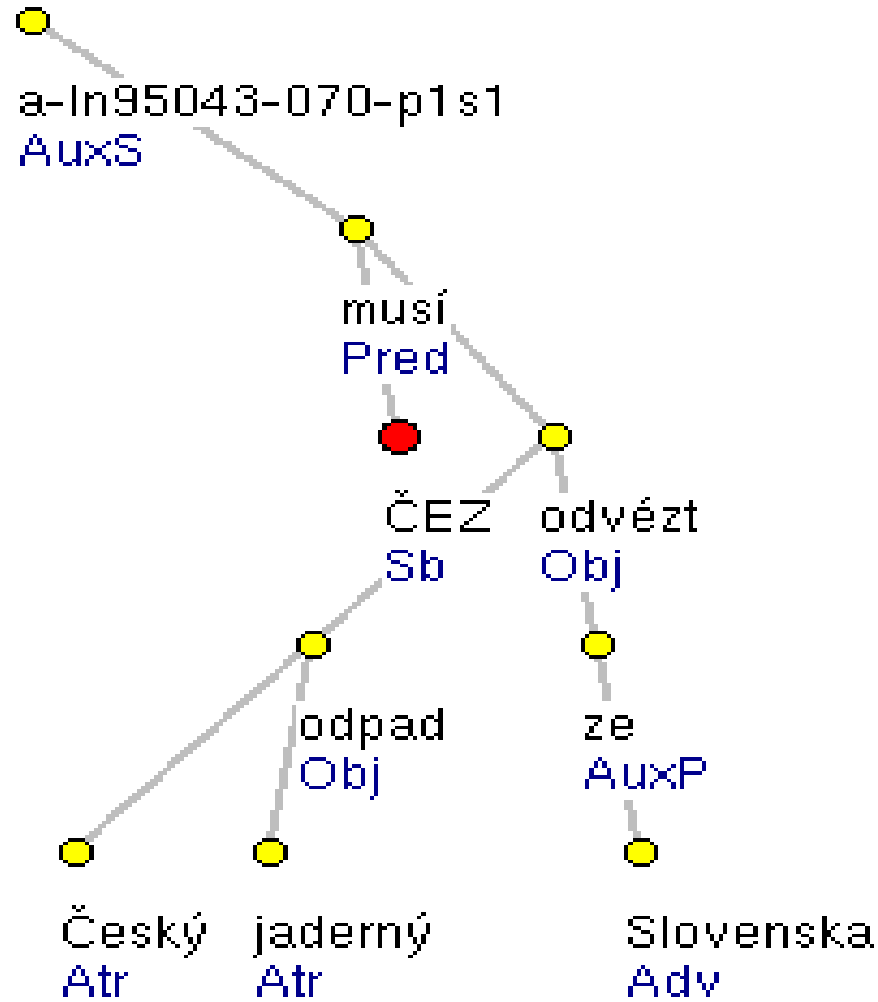
Data v PDT2

- závislost (a/t rovina)
 - jednoduchý, čistý a intuitivní koncept
- a/t stromy ale nepopisují jen závislost
 - koordinace
 - elipsy
- důsledek – fráze nemusí být (pod)stromem
- kombinace jevů – nepřehledná struktura
 - nevhodné pro automatické zpracování

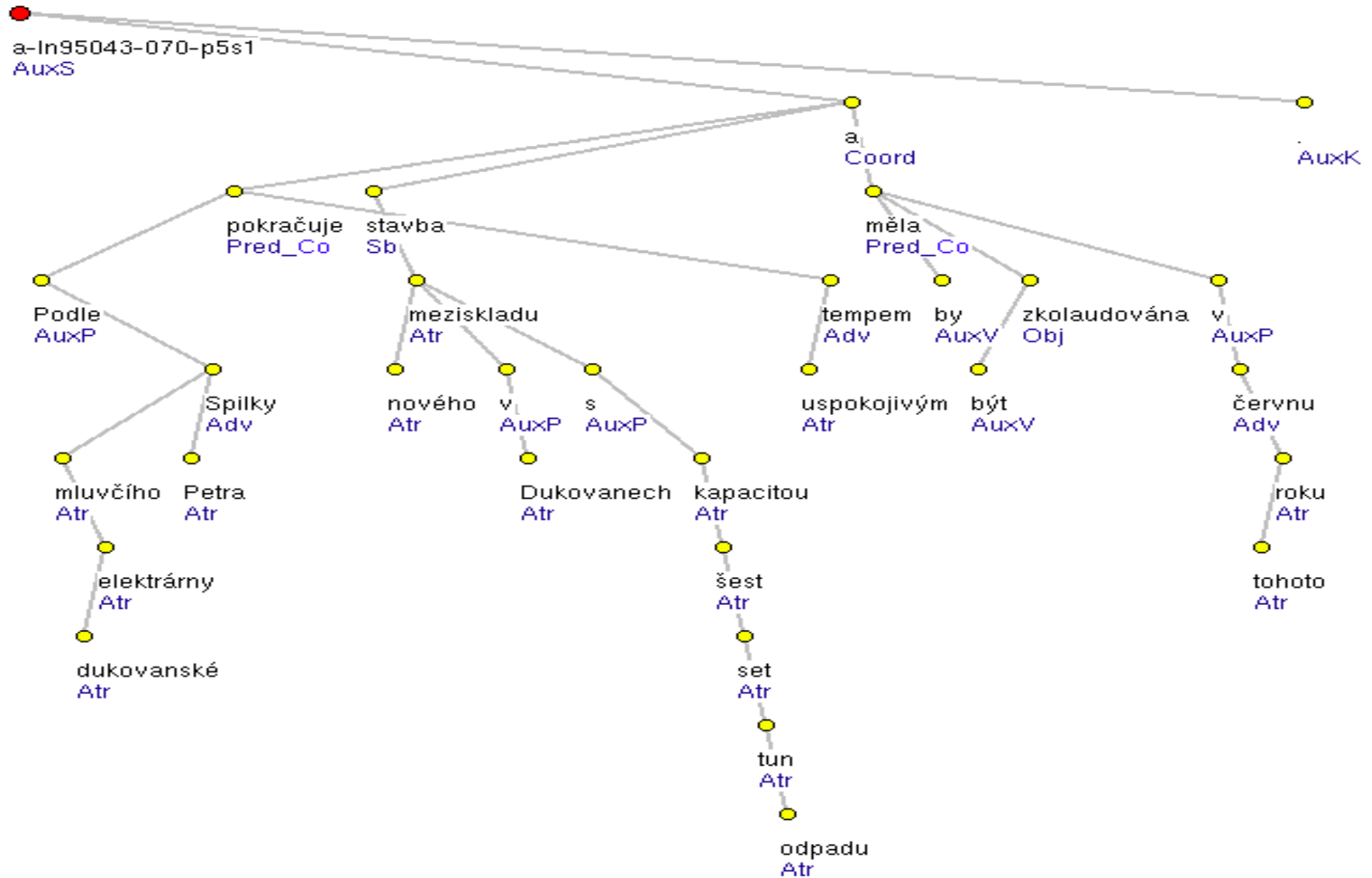
Strom, a-rovina



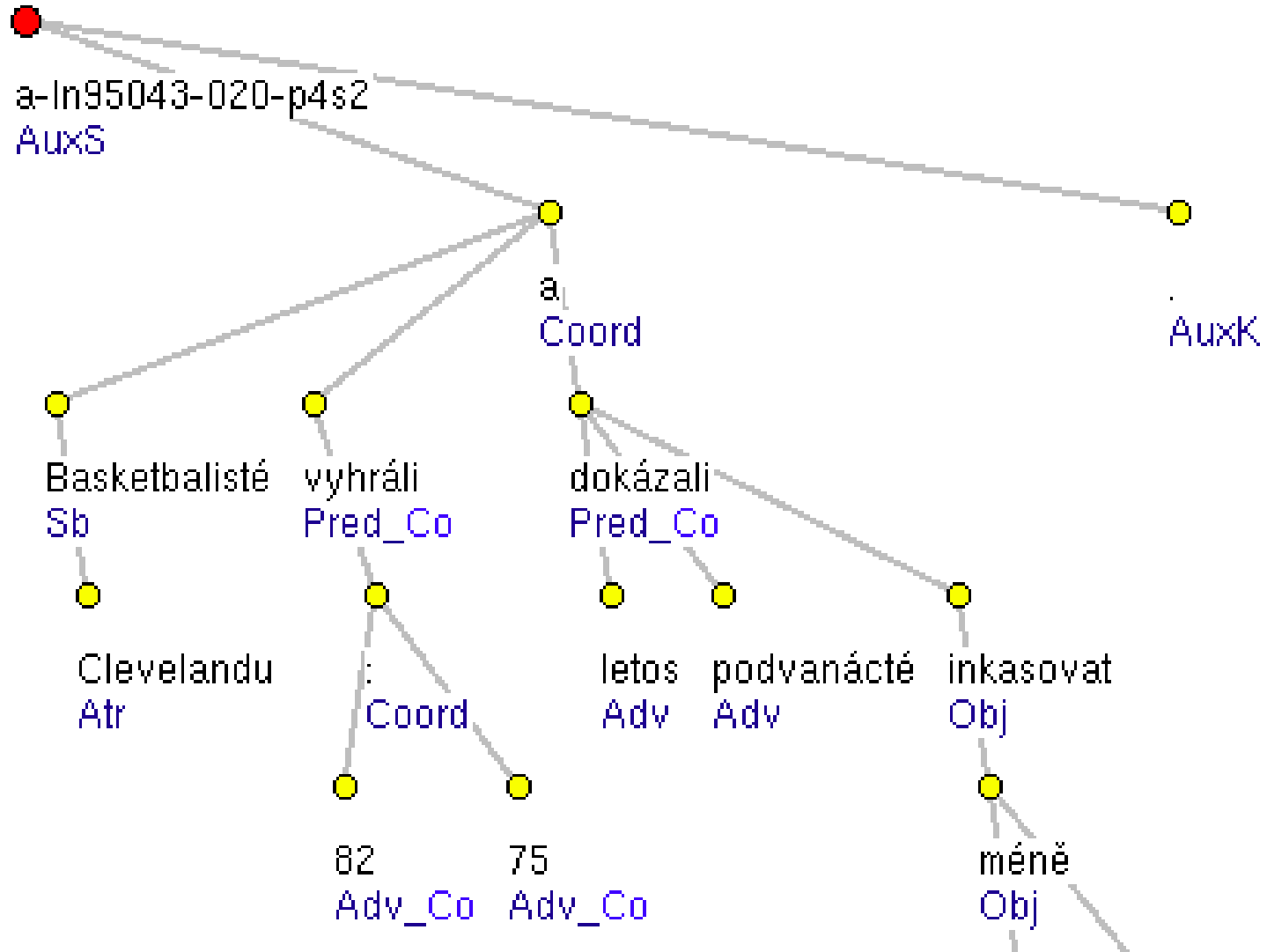
A-rovina



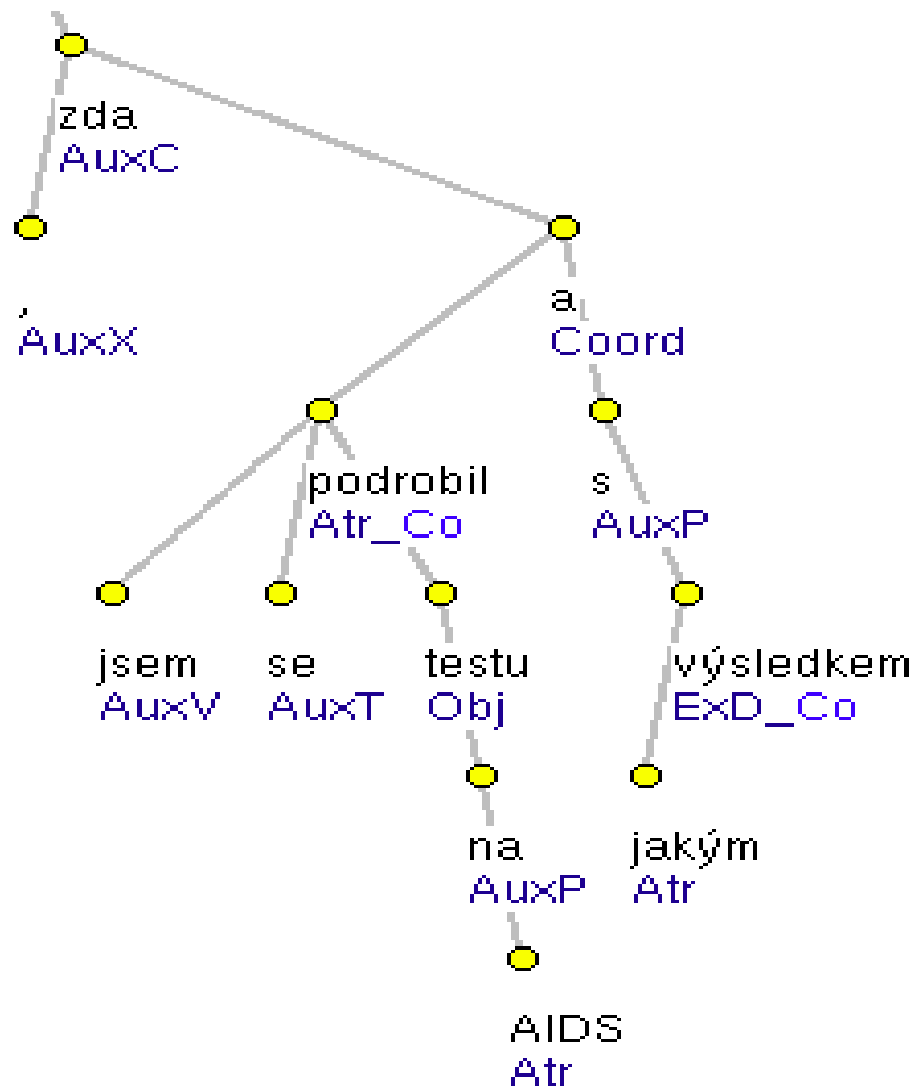
A-rovina



A-rovina, koordinace

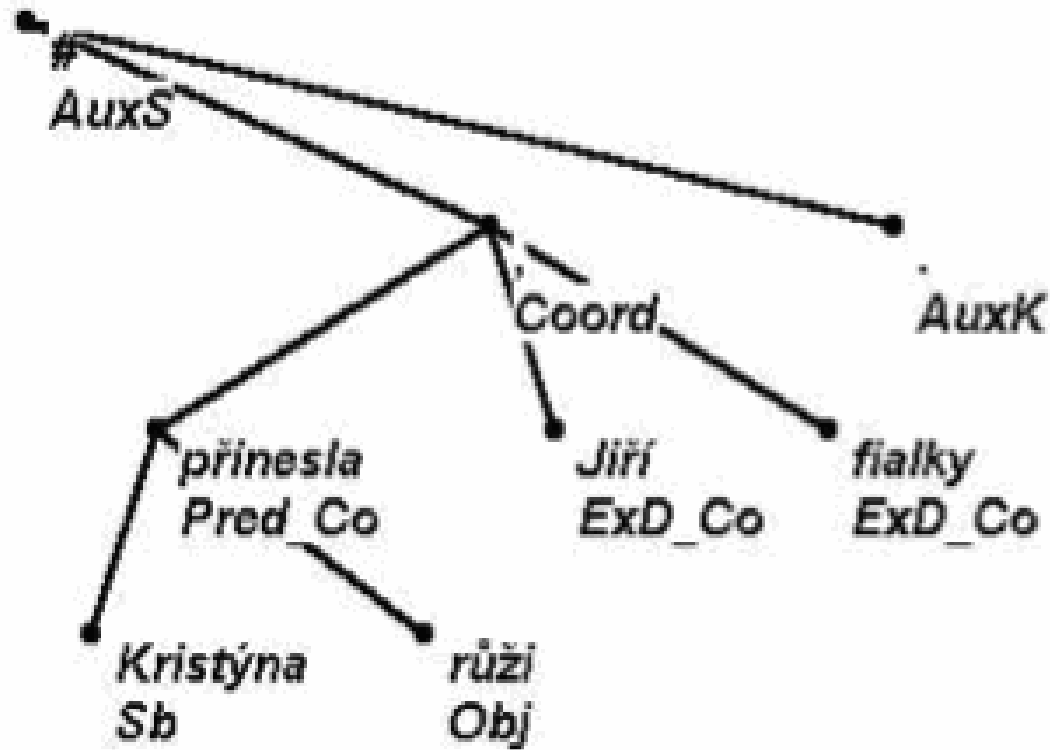


A-rovina, smíšená koordinace



A-rovina, smíšená koordinace

- koordinace klause a PP



Teorie vs. praxe

- teoreticky čisté a pochopitelné
- prakticky:
- Není spoleh na běžné a jednoduché jevy
 - okrajové jevy mění strukturu a vlastnosti
- Co se naučí strojové učení?
 - *technické* zachycení okrajových jevů
 - netextové *smetí*, hrana za každou cenu (adresy, aritmetika, bibliografické záznamy)

Klause

- Manuál udává pouze teoretickou definici
- Formalismus klause nijak nezachycuje
- Je téměř nemožné je detekovat spolehlivě
- poměrně důležitá syntaktická/sémantická jednotka
- patrně podkladem pro některá rozhodnutí anotátorů ...

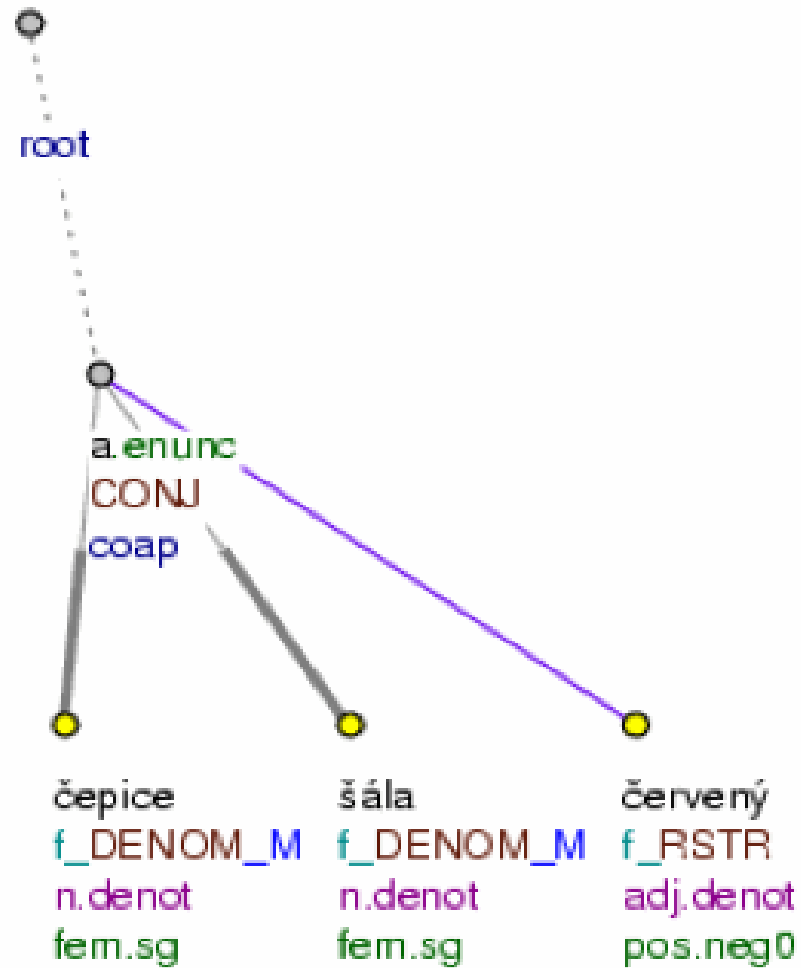
Klause

- klause = finitní VP a její aktanty
- heuristická procedurální detekce
- finitní VP jako oddělovače
 - její sbírání z okolí ve stromu
- koordinace
 - slučování s nefinitními frázemi
- procedurální + úspěšnost <100% (!!!)

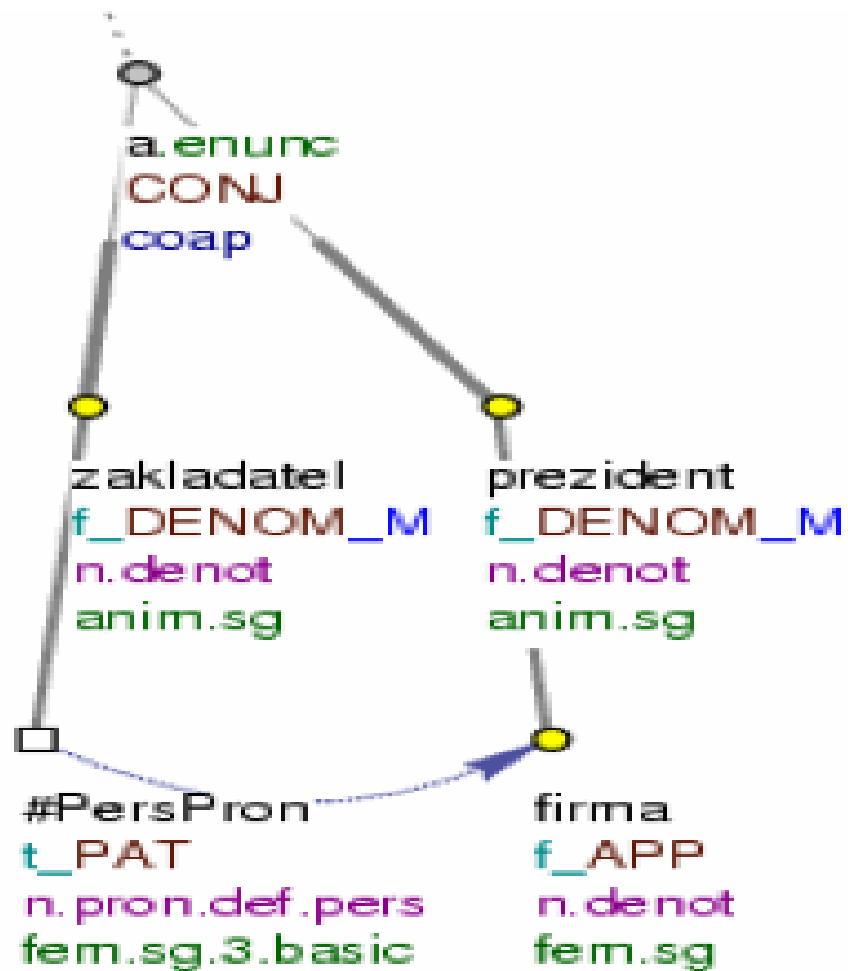
T-rovina

- hloubková syntax, neboli napůl sémantika
- zjednodušení stromové struktury
 - pouze základní sémantické vztahy
- naopak, doplnění nevyjádřených uzlů
- zachycení mnoha jevů
 - mnoho pravidel, mnoho výjimek

T-rovina, koordinace



“zakladatel a prezident firmy”



Anotační manuál

- A-rovina: 301 stran
- T-rovina: 1215 stran (!!!)
- mnoho velmi jemných rozlišení:
 - #Gen – **general actor**
“Domy se stavějí z cihel.”
 - #Unsp – **unspecified actor**
“Na poště zavírají v šest.”
- mnohá mění strukturu stromu
- mnohá lze považovat poněkud vágní

Teorie vs. praxe

- \approx zajímavé vs. užitečné
- Užitečnost PDT by podle mne vzrostla zjednodušením anotace.
- Složité jevy působí téměř jako šum.
- Nicméně: mnoho velmi cenných dat
- nabízející se východisko:
 - co nejlepší ad hoc extrakce pro daný účel
- v mém případě: klause, NP, zájmena
 - vertikál s jednoduchými tagy (btred?)

Děkuji za pozornost.