Pražský závislostní korpus 2.0

Jiří Materna

4. března 2007

Co je PDT 2.0

- Ručně anotovaný korpus českých textů
- Značky postihují morfologii, syntax i sémantiku
- Následník PDT 1.0
- Celkem 2 miliony slov
- 2 mil. morfologie, 1.5 mil. syntax, 0.8 mil. sémantika

Určení PDT 2.0

- Explicitně ověřit teorii funkčně generativním popisu (FGD)
  - použití závislostní syntaxe
  - zahrnutí hloubkové syntaktické roviny do lingvistického popisu
  - formální popis informační struktury věty
- Automatická analýza českých vět
- Generování správně utvořených českých vět
- Strojový překlad

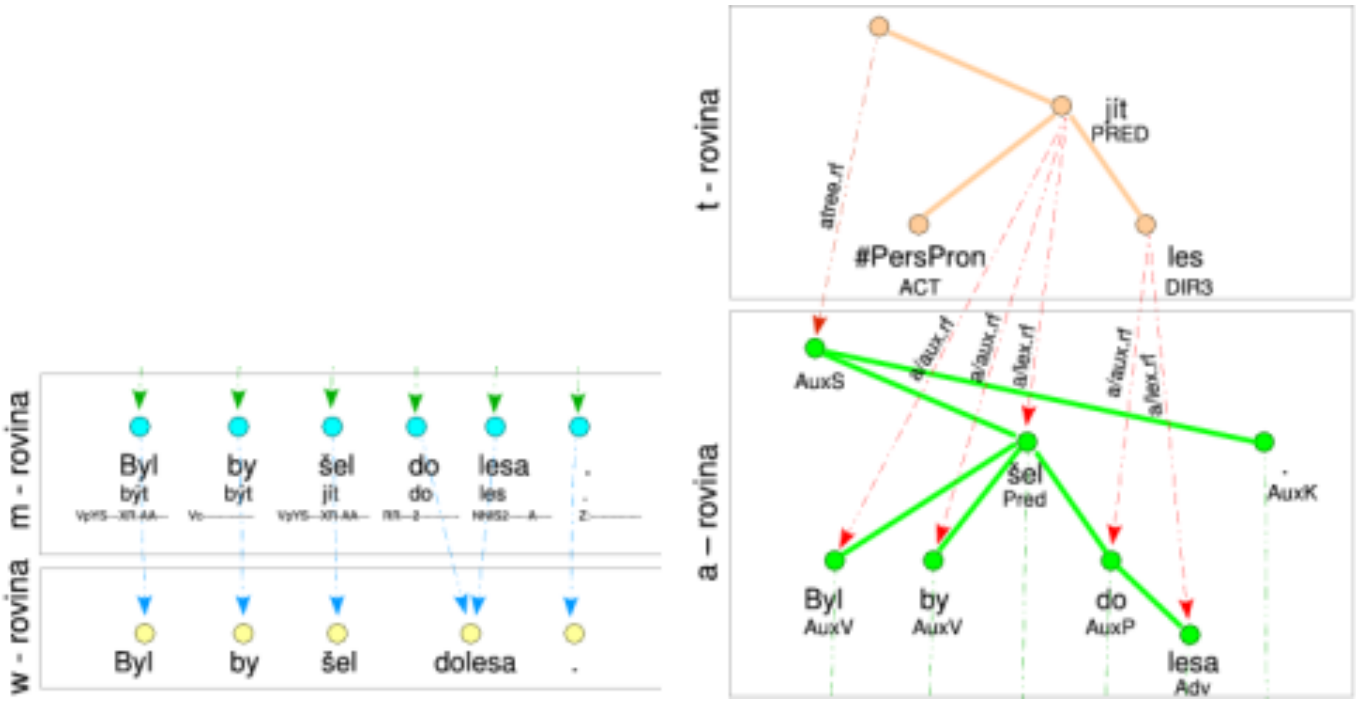
Historie PDT

- Pražský lingvistický kroužek
- Inspirací anglický Penn Treebank
- Nejprve pouze morfologická rovina
- Po doplnění syntaktických závislostí vznikl PDT 1.0
- V roce 2006 zveřejněna verze PDT 2.0, obsahující i sémantické informace

Roviny anotace

- Slovní rovina
- Morfologická rovina
- Analytická rovina
- Tektogramatická rovina
- Všechny roviny jsou navzájem propojené

Propojení rovin



Morfologická rovina

- Rozdělení textu do vět a přiřazení atributů slovním jednotkám
- Pro každé slovo 3 hlavní atributy
  - *lemma*
  - *tag* – 15 pozic
  - *id* – jednoznačná identifikace slova ve větě
- opravný atribut *form*

Morfologická rovina – anotace

- Nejprve předzpracování morfologickým analyzátorem
- Opravení chyb dvěma nezávislými anotátory
- Neshody anotátorů opraveny třetím anotátorem

Analytická rovina

- Reprezentována orientovaným závislostním stromem
- Každému uzlu odpovídá slovní jednotka z morfologické roviny
- Hrany jsou ohodnoceny svým typem:
  - závislostní vztah
  - apozice
  - koordinace
  - apod.
- Zaznamenáno pořadí slov ve větě
- každý uzel má 6 atributů (*id, ord, m.rf, . . .*)

Analytická rovina – anotace

- V první fázi úplná ruční anotace
- Na hotových datech naučen parser
- Parser spuštěn na zbytku dat
- Data opravena anotátory
- Neshody anotátorů rozhodnuty dalším anotátorem

Tektogramatická rovina

- Reprezentována orientovaným závislostním stromem
- Uzly zastupují pouze plnovýznamová slova
- Ne všechny prvky morfologické roviny musí být zastoupeny, nebo naopak (nevyjádřený podmět)
- Aktuální členění (TFA, Topic-focus articulation)
  - kontextově zapojený
  - kontrastivně kontextově zapojený
  - kontextově nezapojený
- Vyznačení koreference

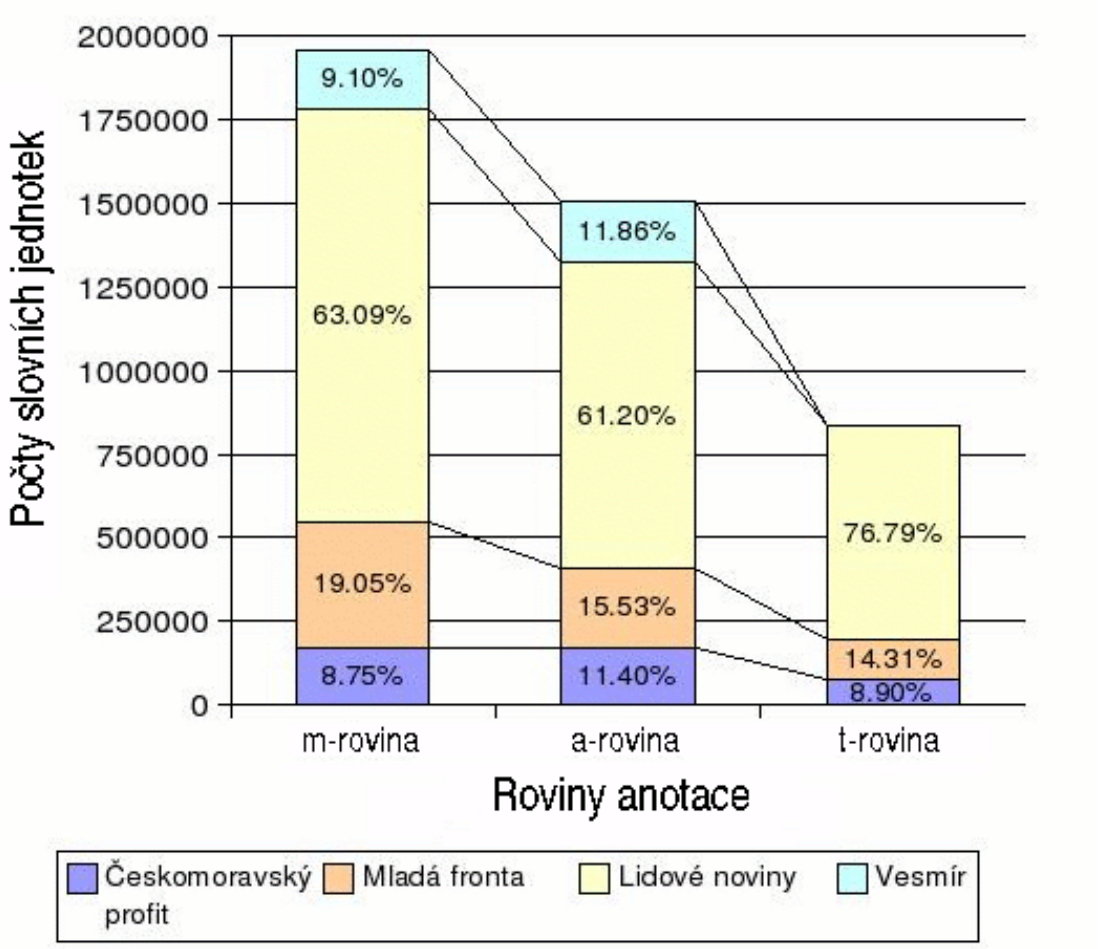
Tektogramatická rovina – anotace

- Stejný postup jako u analytické roviny
- Aktuální členění, koreference a jiné atributy anotovány ručně
- Na závěr kontrola propojení mezi všemi rovinami

Zdroje dat

- Textová data čerpána z:
  - Lidové noviny (deník), 1991, 1994, 1995
  - Mladá fronta Dnes (deník), 1992
  - Českomoravský Profit (ekonomický týdeník), 1994
  - Vesmír (populárně vědecký měsíčník), 1992, 1993
- Odstraněny přepisy šachových partií, tabulky výsledků sportovních utkání apod.

Rozdělení dat



Trénovací a testovací data

- Data podle druhu učení rozdělena do 3 skupin
- trénovací data (train)
- vývojová testovací data (test)
- evaluační testovací data (etest)
- poměr rozdělení **train/test/etest** je přibližně **8:1:1**

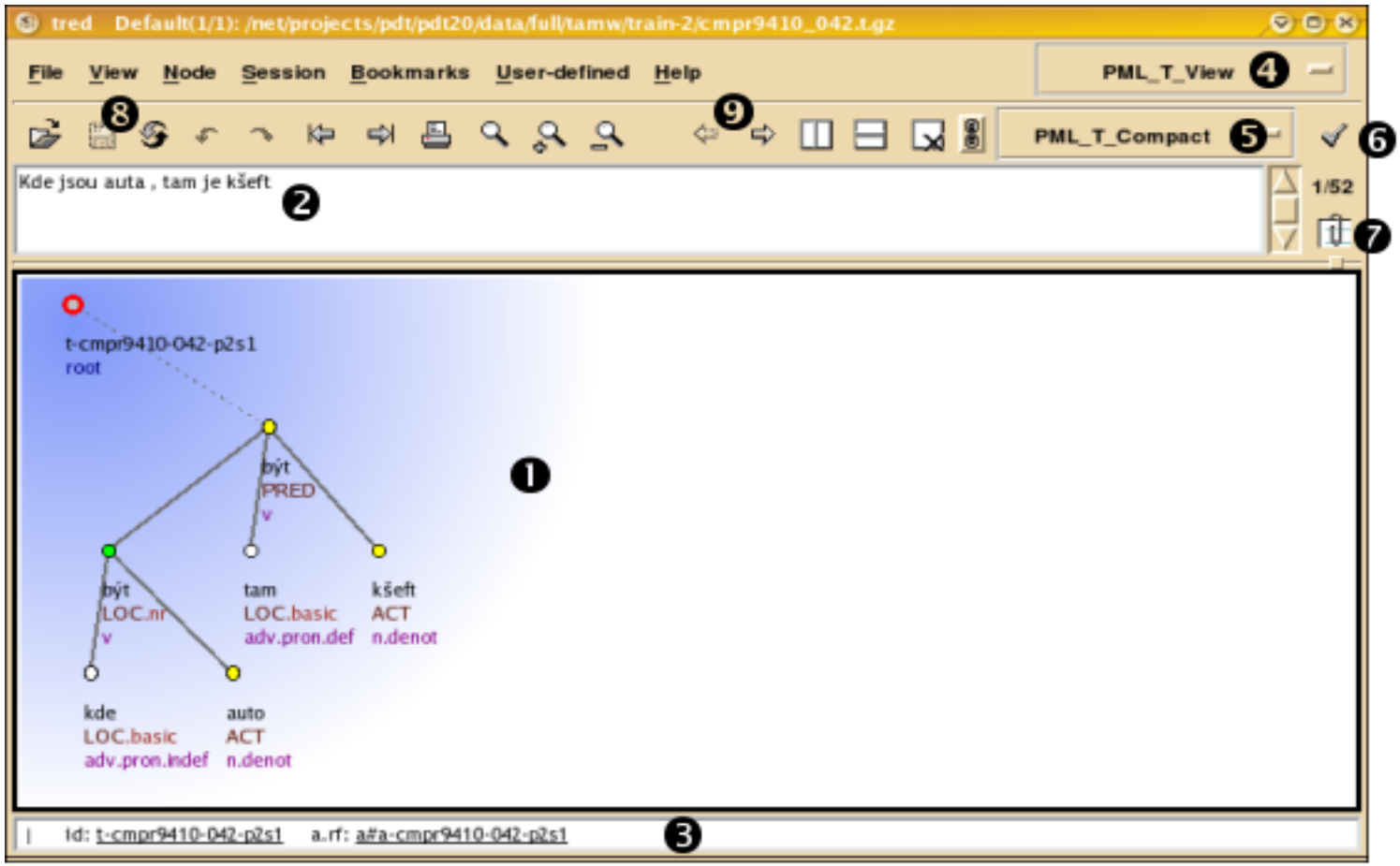
Nástroje – NetGraph

- Aplikace pro prohledávání korpusu
- Klient-server – současné prohledávání více uživateli
- Server napsán v C/C++, klient v Javě
- Dotazem je uzel nebo strom
- Výsledkem dotazu všechny stromy, které zadaný strom obsahují jako podstrom

Další Nástroje

- **TrEd** – editace dat v korpusu
- **Btred**/**Ntred** – perlovské skripty pro přímý přístup k datům
- Dále několik prográmků pro konverzi dat mezi verzemi PDT

TrEd


tred Default(1/1): /net/projects/pdt/pdt20/data/full/tamw/train-2/cmpr9410_042.t.gz
File View Node Session Bookmarks User-defined Help
PML_T_View 4
PML_T_Compact 5
6
8
9
Kde jsou auta , tam je kšeft 2
1/52
7
t-cmpr9410-042-p2s1
root
být
PRED
v
1
být
LOC.nr
v
tam
LOC.basic
adv.pron.def
kšeft
ACT
n.denot
kde
LOC.basic
adv.pron.indef
auto
ACT
n.denot
id: t-cmpr9410-042-p2s1 a.rf: a#a-cmpr9410-042-p2s1 3

Budoucnost PDT 2.0

- Přidání mluvených dat
- Přidání hlubší a širší anotace obzvláště pro koreferenci
- Anotace jiného odlišného jazyka, např. angličtiny
- přidání dalších vrstev anotace (reprezentace znalostí založená na obsahu výpovědi)