Snídaně s Českým akademickým korpusem (informační schůzka)

25. června 2007, 8:45

zadání:
Anotování Českého akademického korpusu

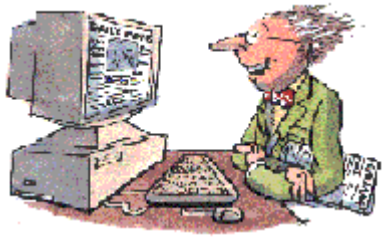
- Co to znamená?

Korpus

- textový korpus … banka textů
- anotovaný korpus … texty obohaceny lingvistickou informací (např. slovní druhy, morfologické kategorie, syntaktický rozbor, …)

Proč anotování korpusů?

- teoretická lingvistika
  - materiál k teoretickému bádání
- počítačová lingvistika
  - strojové učení
  - trénovací data: anotované korpusy (čím více, tím lépe – i když všechno má svoji mez)
  - aplikace, např.
    - tagging – automatické určování slovních druhů a morf. kategorií
    - parsing – automatické provádění syntaktického rozboru



"The computer is claiming its intelligence is real, and ours is artificial."

Navštivte stránku

popularizačních článků a rozhovorů

http://ufal.mff.cuni.cz/?a=popular&m=student_info

projekt „Data a nástroje pro informační systémy" (REST)

- http://ufal.mff.cuni.cz/rest
- datová komponenta – Český akademický korpus (ČAK)
- komponenta nástrojů – nástroje pro morfologické zpracování textů (anotace, morfologická analýza, tagging)

Český akademický korpus

- 1971–1985 – ÚJČ AV
- 540 000 morfologicky a syntakticky anotovaných slov
- psané a mluvené texty
- publicistika, administrativa, odborné statě
- anomálie
  - vymazaná interpunkce
  - vymazané ciferné výrazy

OJEDINĚLÝ

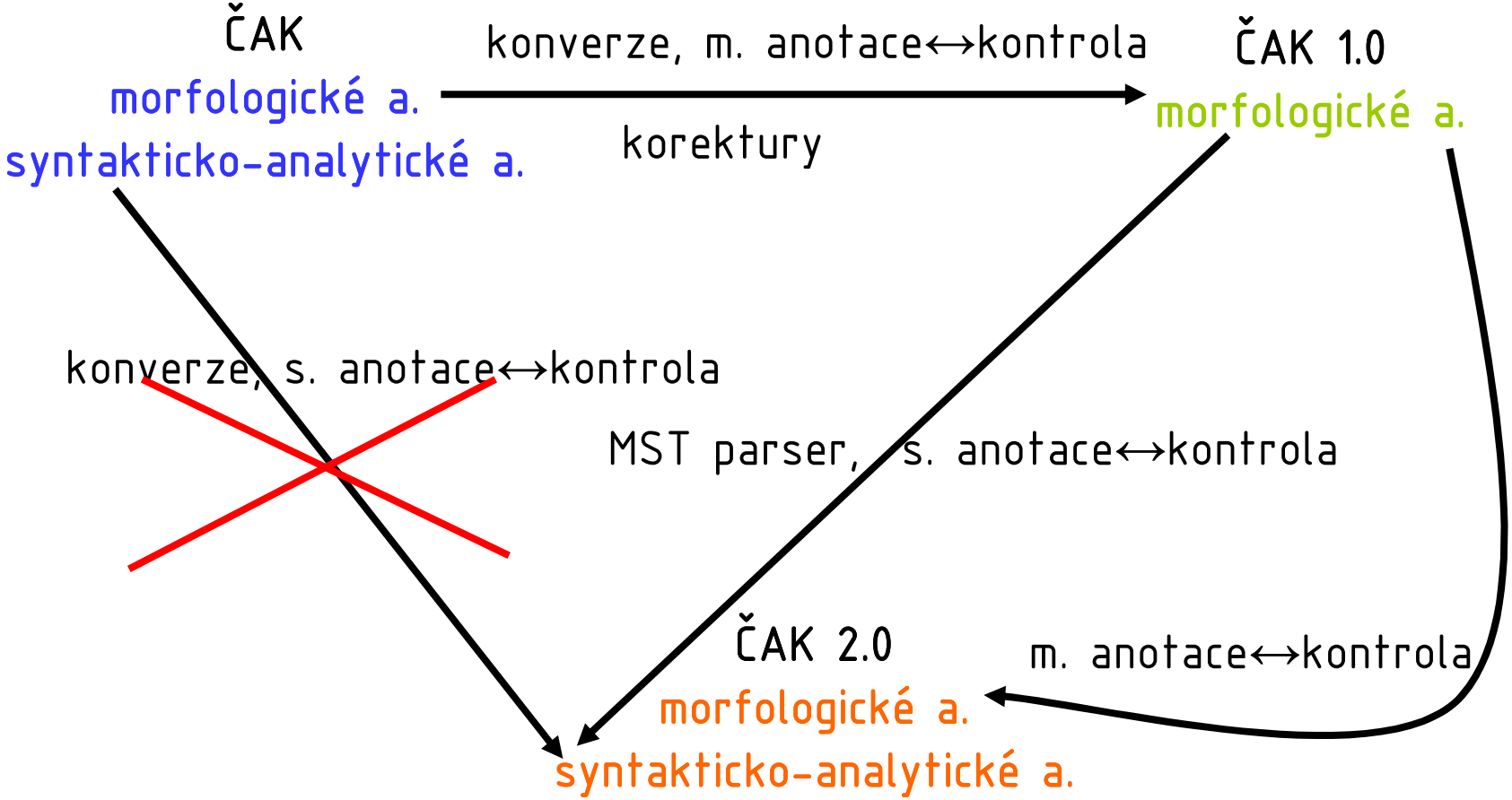
Proč jsme si na něj vzpomněli?

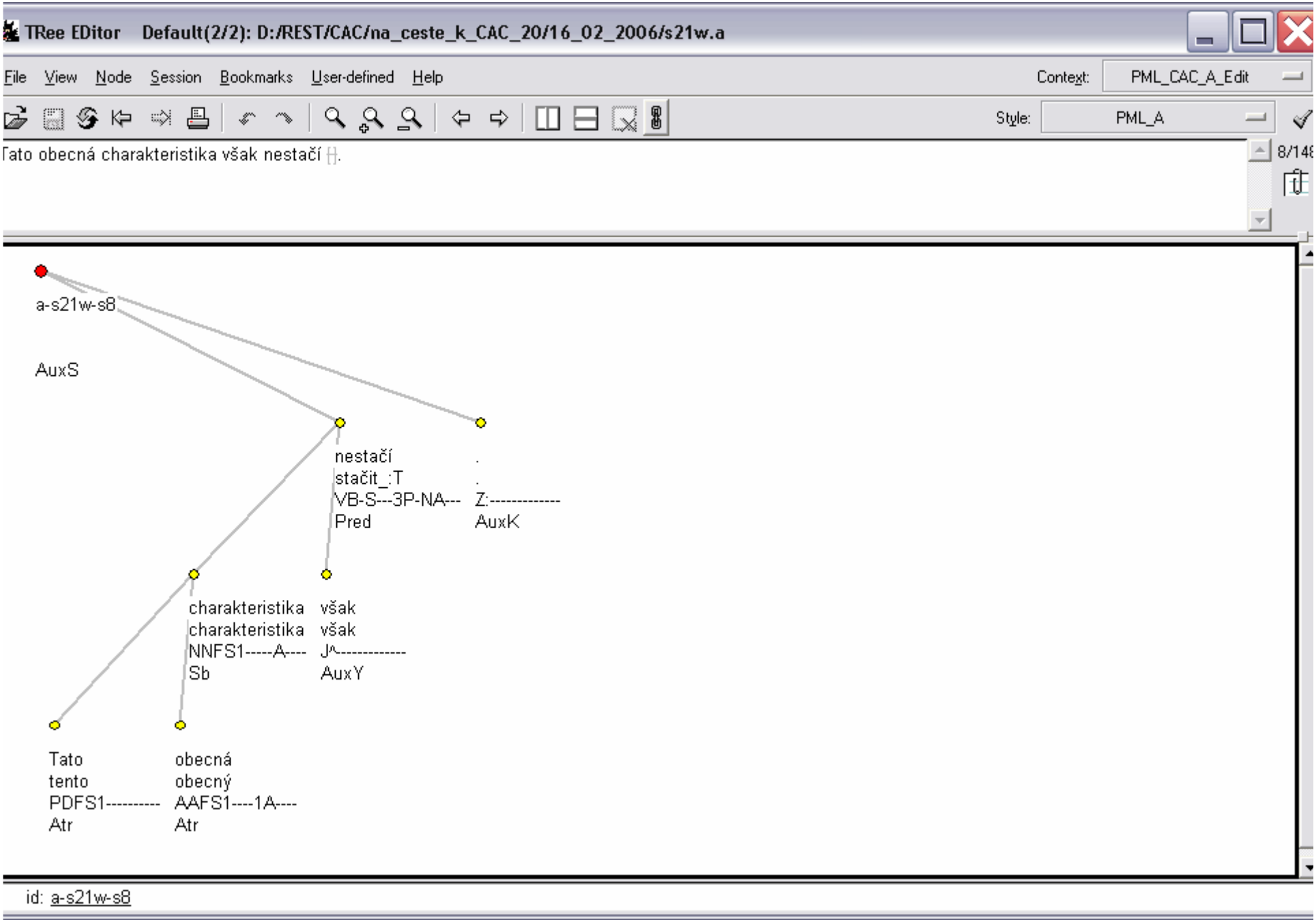
- 1994 – první experimenty strojového učení na češtině, konkrétně tagging a ČAK jako trénovací data. Zásadní z pohledu dalšího vývoje komputační lingvistiky v Čechách
- 1996–2006 anotace Pražského závislostního korpusu (PZK, 2006 vydaná již druhá verze)
  - http://ufal.mff.cuni.cz/pdt2.0
  - anotace morfologické, syntaktické, tektogramatické

Proč jsme si na něj vzpomněli? (2)

- Nápad: obohatit 80 000 syntakticky anotovaných vět z PZK o 30 000 vět z ČAK
- Odlišnosti PZK vs. ČAK
  - vnitřní formát (připomeňte si, kdy ČAK vznikl)
  - anotační schémata
  - chybějící interpunkce a ciferné výrazy
- Proto konverze ČAK do „podoby" PZK
  - třeba to půjde úplně automaticky …

Nepůjde.


ČAK
morfologické a.
syntakticko-analytické a.
konverze, m. anotace↔kontrola
korektury
ČAK 1.0
morfologické a.
konverze, s. anotace↔kontrola
MST parser, s. anotace↔kontrola
ČAK 2.0
morfologické a.
syntakticko-analytické a.
m. anotace↔kontrola


TRee EDitor Default(2/2): D:/REST/CAC/na_ceste_k_CAC_20/16_02_2006/s21w.a
File View Node Session Bookmarks User-defined Help
Context: PML_CAC_A_Edit
Style: PML_A
Tato obecná charakteristika však nestačí .
8/148
a-s21w-s8
AuxS
nestačí
stačit_:T
VB-S---3P-NA---
Pred
.
.
Z:----------
AuxK
charakteristika
charakteristika
NNFS1-----A----
Sb
však
však
J^----------
AuxY
Tato
tento
PDFS1----------
Atr
obecná
obecný
AAFS1----1A----
Atr
id: a-s21w-s8

Na cestě k syntakticko-analytickým a. ČAK 2.0

- MST parser

  VSTUP: ČAK 1.0

  VÝSTUP: automatický syntaktický rozbor

  ÚSPĚŠNOST: 84,6% (na hranách)

- s. anotace↔kontrola

  VSTUP: automatický syntaktický rozbor

  VÝSTUP: ručně opravené chyby (chybně určené závislosti a analytické funkce) automatické procedury

s. anotace↔kontrola

viz zadání na začátku

1. dvojitá anotace v editoru TrEd (ar[illegible])
2. technická podpora v TrEd pro řešení mezianotátorských odlišností; (ne)shody v číslech (Kiril Ribarov)
3. vyřešení odlišností v TrEd (Alla Bémová, Zdeňka Urešová)
4. automatické kontrolní skripty àla kontrola PDT 2.0 (Jiří Mírovský)

paralelně m. anotace ↔ kontrola (Jiří Mírovský)

… už by mělo být jasné, o co v brigádě půjde …

Anotace ČAK technicky

- podrobně rozepsáno v dokumentu AnalAnnotCAC_07.pdf

AnalAnnotCAC_07.pdf

1. Motivace
2. Český akademický korpus
3. Anotace jako zadání
4. Anotace technicky
5. Pokyny k anotaci
6. Anotační nástroj TrEd
7. Zahřívací kolo
8. Meetpoint

Kolik času?

- zaučení – červenec (horní odhad)
- anotace – srpen, září, říjen
- zkušenosti – max 100 vět denně za cca 5h
- ≈ čtyřměsíční poloviční úvazek
- Rozmyslete si prosím velmi dobře, jestli budete mít dost času …

Počítače

- můžeme zapůjčit notebooky

Odměna

- 5 Kč à věta + odměny
- DP[ČP]

Co dál?

- vážní zájemci projdou zahřívacím kolem
- vybereme 4 (příp. 5), kteří udělají nejmenší počet chyb
- soubory budou předávány i odevzdávány elektronicky
- přítomnost na ÚFAL není nutná