Snídaně
s Českým akademickým korpusem

18. června 2007, 8:30

Anotování
Českého akademického korpusu

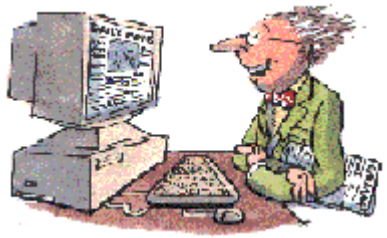
- PROČ? JAK?

Kdo

- Hosté
  - Veronika Čurdová
  - Lenka Žehrová
- Domácí … viz dále

Anotování korpusu. Proč?

- teoretická lingvistika



- počítačová lingvistika
  - strojové učení
  - trénovací data: anotované korpusu
  - aplikace: tagging, parsing, ...

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

Navštivte stránku

popularizačních článků a rozhovorů

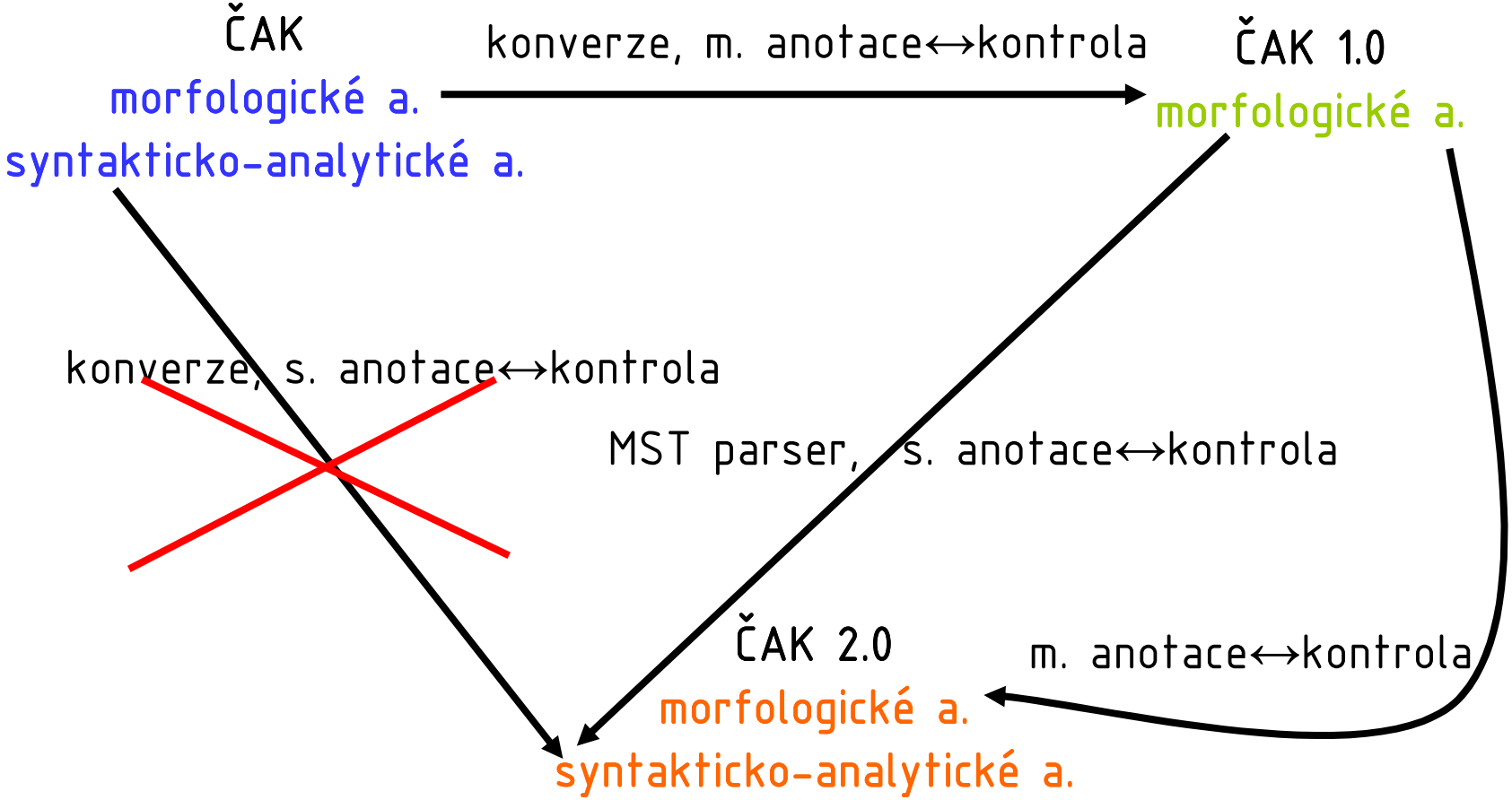
http://ufal.mff.cuni.cz/?a=popular&m=student_info

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

- http://ufal.mff.cuni.cz/rest
- datová komponenta – Český akademický korpus
- komponenta nástrojů

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


ČAK
morfologické a.
syntakticko-analytické a.
konverze, m. anotace↔kontrola
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

Na cestě k syntakticko-analytickým a. ČAK 2.0

- MST parser

  VSTUP: ČAK 1.0 (w, m soubory)

  VÝSTUP: a' soubory

  ÚSPĚŠNOST: 84,6% (na hranách)

- s. anotace↔kontrola

  VSTUP: a' soubory

  VÝSTUP: a soubory

s. anotace↔kontrola

1. dvojitá anotace v editoru TrEd (anotátoři)
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)

s. anotace↔kontrola: TrEd

Zvláštní pozornost:

- Správné použití tzv. **kontextů** v TrEdu. Nově upravený kontext pro snadnou anotaci PML_CAC_A_Edit
- **guessed_form** je součástí m-souboru. Slouží k upřesnění, dle chápání anotátora, obsahu prázdných uzlů (původně chybějících slov, či někdy špatně vložených slov).
  Pozor: Nutně ukládat i m-soubor -> viz „Save"
- Při anotaci **koordinace** a apozice, nastavit u všech relevantních uzlů podstromu hodnotu is_member na 1.
- Při anotaci **parentéze**, nastavit hodnotu u všech relevantních uzlů podstromu hodnotu is_parenthesis_root na 1.

s. anotace↔kontrola: od anotátorů

- m-soubor (původní název)
- a-soubor (původní název doplněný iniciálami)
- poznámky

Termín konce …

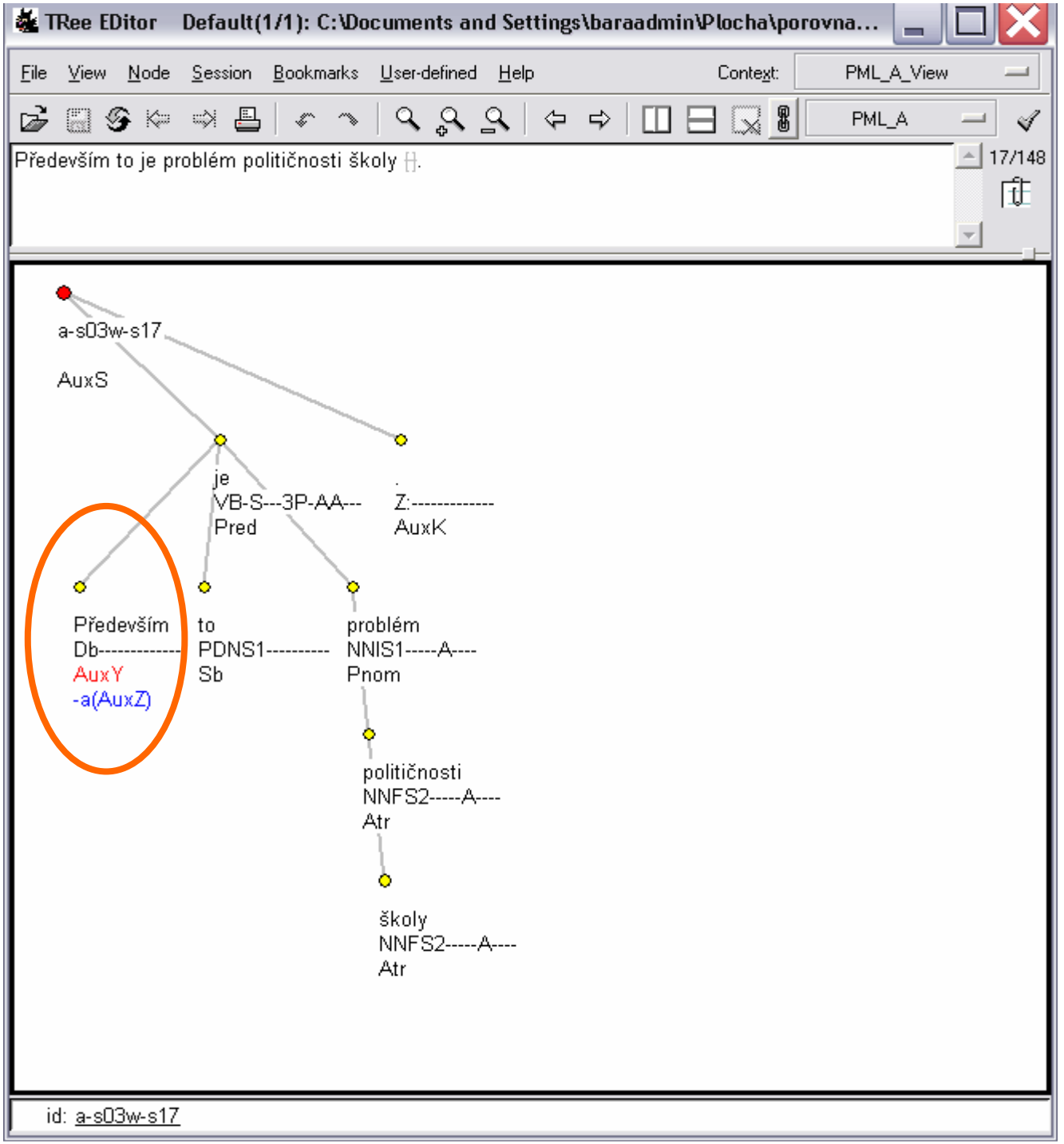
- nesděluji. Stačí, když jsem nervózní já, příp. pane ředitel☺
- Při dodržování dílčích termínů vše stihneme v termínu!

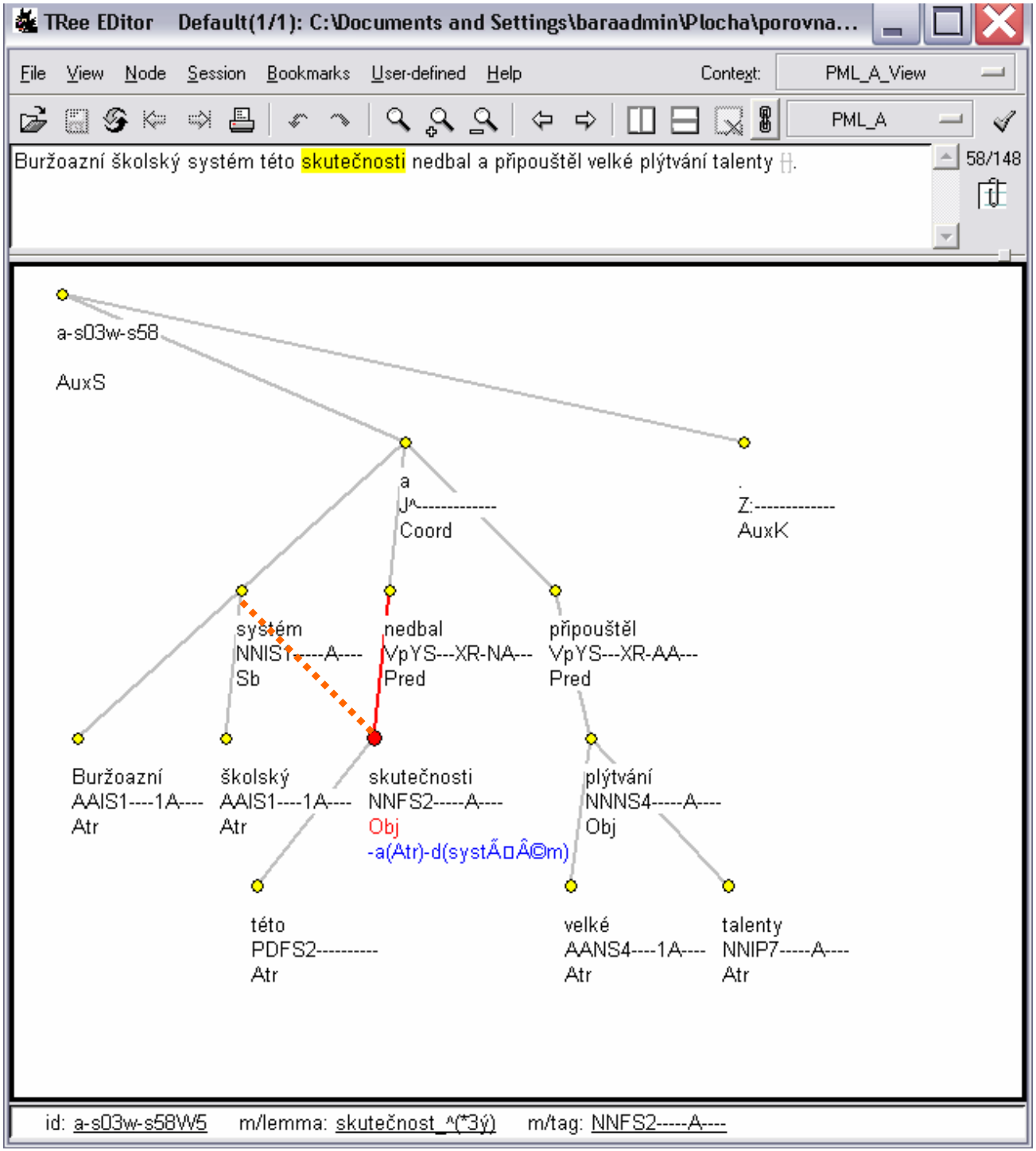
Odměna

- 5 Kč à věta + odměny
- DP[ČP]

Co dál?

- demonstrace TrEd
- debata nad odlišnostmi v anotacích


TRee EDitor Default(1/1): C:\Documents and Settings\baraadmin\Plocha\porovna...
File View Node Session Bookmarks User-defined Help
Context: PML_A_View
PML_A
Především to je problém političnosti školy .
17/148
a-s03w-s17
AuxS
je
VB-S---3P-AA---
Pred
.
Z:----------
AuxK
Především
Db----------
AuxY
-a(AuxZ)
to
PDNS1----------
Sb
problém
NNIS1-----A----
Pnom
političnosti
NNFS2-----A----
Atr
školy
NNFS2-----A----
Atr
id: a-s03w-s17


TRee EDitor Default(1/1): C:\Documents and Settings\baraadmin\Plocha\porovna...
File View Node Session Bookmarks User-defined Help
Context: PML_A_View
PML_A
Buržoazní školský systém této skutečnosti nedbal a připouštěl velké plýtvání talenty .
58/148
a-s03w-s58
AuxS
a
J^----------
Coord
.
Z:----------
AuxK
systém
NNIS1----A----
Sb
nedbal
VpYS---XR-NA---
Pred
připouštěl
VpYS---XR-AA---
Pred
Buržoazní
AAIS1----1A----
Atr
školský
AAIS1----1A----
Atr
skutečnosti
NNFS2-----A----
Obj
-a(Atr)-d(systÃ¤Â©m)
plýtvání
NNNS4-----A----
Obj
této
PDFS2---------
Atr
velké
AANS4----1A----
Atr
talenty
NNIP7-----A----
Atr
id: a-s03w-s58W5 m/lemma: skutečnost_^(*3ý) m/tag: NNFS2-----A----