Rukověť anotátora Českého akademického korpusu (transformace syntakticko-analytických anotací)

Barbora Hladká (hladka@ufal.mff.cuni.cz)

20. dubna 2007, 23. června 2007

Obsah

1, Motivace

Český akademický korpus (ČAK) vznikl před více jak dvaceti lety a ve své době byl svým objemem a svými anotacemi ojedinělý (bohužel politické klima té doby nebylo nakloněno jakémukoli šíření slávy za hranice země české) - 550 tis. slov bylo anotováno na morfologické a syntakticko-analytické rovině. Podrobně je historie ČAK dokumentována v práci (Hladká, Králík, 2006).

Vzhledem k nezanedbatelnému objemu a vzhledem ke zkušenostem získaným během práce na Pražském závislostním korpusu jsme se rozhodli převést vnitřní formát a anotační schémata ČAK do formátu a schémat PZK tak, aby ČAK mohl být začleněn do některé z budoucích verzí PZK. Vydáním ČAK 1.0 jsme ukončili první etapu převodu, a to převodu vnitřního formátu a morfologických anotací – podrobné informace viz http://ufal.mff.cuni.cz/rest a (Hladká a kol., 2007; Hladká, Králík, 2006).

Nyní je na řadě převod syntakticko-analytických anotací, který již začal pilotní studií. Studie měla ukázat, odkud a jak se na převod vrhneme. Průběh a závěry studie jsou shrnuty v práci (Ribarov, Bémova, Hladká, 2006). Studie ukázala, že ruční zásah bude nutný (jak se dalo očekávat). Proto se další odstavce budou týkat onoho ručního zásahu, o kterém nadále mluvíme jako o anotaci.

2, Český akademický korpus

ČAK obsahuje dokumenty tří stylů (publicistický, administrativní, odborný) a dvou forem (psaná, mluvená[1]). Následující tabulka shrnuje číselné charakteristiky ČAK. Sloupec nadepsaný titulkem počet „zvětšených“ vět potřebuje zvláštní komentář, protože zosobňuje jednu výraznou anomálii ČAK.

STYL	FORMA	POČET DOKUMENTŮ	POČET VĚT	POČET SLOV	POČET „ZVĚTŠENÝCH“VĚT (%)
publicistický	psaná	52	10 234	189 435	1 563 (15)
publicistický	mluvená	8	1 433	28 737	30 (2)
administrativa	psaná	16	3 362	58 697	915 (27)
administrativa	mluvená	4	989	14 235	15 (2)
odborný	psaná	68	11 113	245 175	2 030 (18)
odborný	mluvená	32	4 576	115 853	113 (2)

[1] texty jsou přepisy mluvených projevů

všechny styly	psaná	136	24 709	493 307	4 508 (18)
všechny styly	mluvená	44	6 998	158 825	158 (2)
všechny styly	psaná a mluvená	180	31 707	652 132	4 668 (15)

Během původních anotací ČAK byla z textu vypuštěna interpunkční znaménka. Zároveň byly vypuštěny ciferné výrazy, tj. číslovky zapsané ciframi (např. 1999). Toto rozhodnutí má svoje opodstatnění – hlavní motivací pro práci na anotovaném korpusu bylo připravit materiál pro kvantitativní studii psané češtiny té doby. Interpunkce a ciferné výrazy se buď nečtou, nebo se mohou číst několika různými způsoby, tedy jsou z pohledu kvantitativní analýzy nezajímavé. Vzhledem k záměru použít data jako trénovací pro metody strojového učení se problém chybějících slovních jednotek ukázal jako závažný. Protože se písemné zdroje textů nedochovaly, bylo nutné data kompletně projít a chybějící jednotky doplnit. Podrobnosti korektur jsou shrnuty na stránce http://ufal.mff.cuni.cz/rest/CAC/CAC_01.html. Zde shrneme základní věci, na které se měli korektoři soustředit:

- Detekovat místa, kde chybí interpunkce (větnou interpunkci bylo možné z původních dat rekonstruovat).
- Detekovat místa, kde chybí ciferný výraz. Kromě detekce místa korektoři specifikovali typ chybějícího výrazu – např. *Akce #A[2] jarních ? bude zahájena #D[3] března*. (n31w, <s id="n31w.m-s100">)
- Upozornit na místa, která se jim nelíbí – podivný pořádek slov ve větě, chybějící slovo – např. *V dalším programu ? Suchomel přednesl hlavní obsah referátu ? Muchy o energetické situaci, která i pro ? je značně napnutá.* (a06w, <s id="a06w.m-s131">)

Vrátíme-li se zpět k předcházející tabulce, počtem „zvětšených" vět myslíme věty, ve kterých korektor zaznamenal druhou a/nebo třetí z uvedených anomálií (dle prezentovaného pořadí). Počtem slov myslíme počet původních slov spolu s vloženou interpunkcí, ciferných výrazy a otazníky (jako se zástupci podivností v textech).

3, Anotace jako zadání

VSTUP: Věty dokumentů ČAK: stromečky s analytickými funkcemi. Stromečky i analytické funkce jsou výsledkem automatické procedury.

VÝSTUP: Výstup automatické procedury není úplně v pořádku, kontrola a oprava stromečků a analytických funkcí je nutná. Anotátor musí zkontrolovat jak stromeček (tedy všechny závislosti), tak i analytické funkce. Výstupem tedy budou opravené stromečky. To vše v souladu s pokyny anotačního manuálu (Hajič a kol., 2004) a s doplňky uvedenými níže.

4, Anotace technicky

- **ČAK ve formátu PML**, tj. soubory `*.a` (syntakticko-analytická rovina) a k nim patřící soubory `*.w` (slovní rovina), `*.m` (morfologická rovina). Ze samotných názvů souborů (viz následující tabulka) je zřejmé, o jaký styl a formu dokumentu se jedná. Druhá a třetí pozice v názvu určují pořadové číslo dokumentu v rámci daného stylu. Např. název `s17w.a` označuje soubor obsahující anotace na analytické rovině odborného psaného dokumentu. Ze souboru vedou odkazy do souboru `s17w.m`.

STYL	FORMA	JMÉNO SOUBORU
publicistický	psaná	n[0-9][0-9]w

[2] množství
[3] datum

publicistický	mluvená	n[0-9][0-9]s
administrativa	psaná	a[0-9][0-9]w
administrativa	mluvená	a[0-9][0-9]s
odborný	psaná	s[0-9][0-9]w
odborný	mluvená	s[0-9][0-9]s

- **TrEd,** jeho klasická distribuce. TrEd je nástroj (stromový editor), pomocí kterého se anotace provádí.

5, Anotační nástroj TrEd

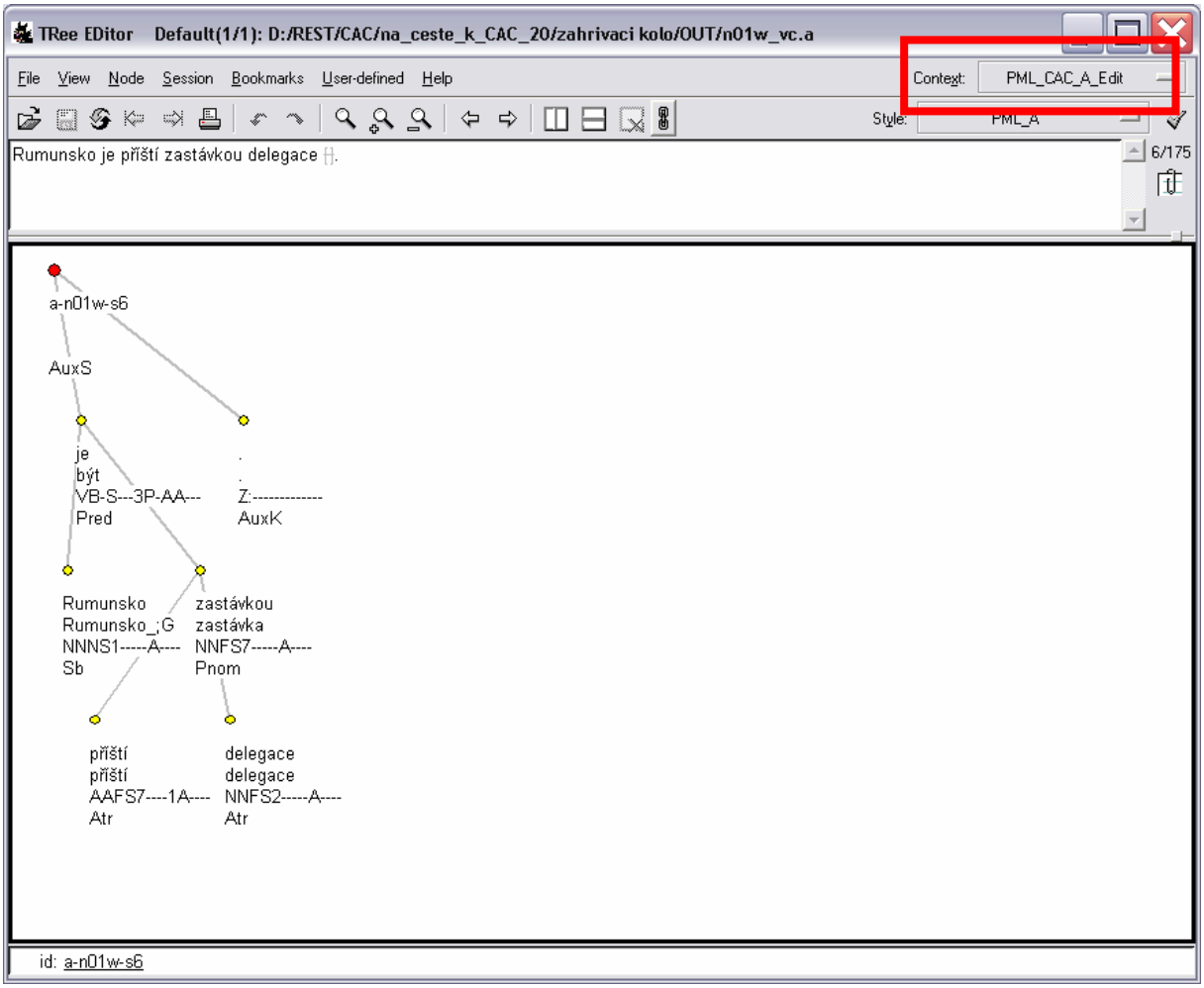
- 19/6/2007 je ke stažení nová verze TrEd, která má v sobě kontext pro anotaci ČAK
  - TrEd je ke stažení ve standardní distribuci (platné i pro anotaci ČAK) na jeho "domácích stránkách" na http://ufal.mff.cuni.cz/~pajas/tred.
  - **Rychlý návod k instalaci** pod systémem Windows (pro "offline" použití; pro anotátory ze SNK instaluje TrEd Radovan jinak):
    - Z výše uvedené stránky stáhněte na svůj notebook soubor `tred_wininst_en.zip`.
    - Rozbalte jej kdekoliv; dostanete adresář `tred_wininst_en`.
    - V tomto adresáři spusťte soubor `setup.bat`. Pozor, v adresáři je několik podobně pojmenovaných `.bat` souborů - vy spusťte tento základní.
    - Na všechny otázky odpovězte yes (y), nebo tak, aby instalace pokračovala, a nechte pro všechno, co vám instalační program nabízí, defaultní hodnoty.
    - Po dokončení instalace spusťte TrEd (pomocí ikony na ploše). Ověřte (Help→About), že se jedná o verzi nejméně 1.3050 nebo vyšší.
  - **Rychlý návod ke spuštění TrEdu** a anotaci (otevření souboru a nastavení prostředí, anotace)
    - Spusťte TrEd, a pomocí `Open` (nebo z `Recent Files`) otevřete příslušný `.a` soubor.
    - Nastavte "kontext" (vpravo nahoře) `PML_CAC_A_Edit`. – viz Obrázek č. 1
    - Používejte makra z `PML_CAC_A_Edit` (`User-defined → PML_CAC_A_Edit`, příp. `More`) – viz Obrázek č. 2 (ze začátku doporučujeme si seznam maker vytisknout (viz bod 8,)) a dále smíte používat `Open`, `Save`, `Save As` z hlavního menu `File`, a všechny další funkce, které slouží k "prohlížení" souboru(ů) - například posouvání po větách, přímý skok na větu s daným pořadovým číslem (tj. různá `GoTo`...), vyhledávání v souboru pomocí F3/F4 apod. **NIKDY** ale nepoužívejte jakékoli funkce, které mění strom nebo hodnoty atributů z menu `Node`, ani z maker `Tred_Macro` nebo jiných maker a kontextů.
    - První uložení souboru po jeho prvním otevření a zahájení jeho anotace
      - Soubor se ukládá pomocí `Save As` (menu `File`), pak zvolte `"Current"` v okně pro volbu formátu.
      - Soubor uložte pod původním jménem s přidaným podtržítkem a iniciálami vašeho jména; postup: zvolte původní soubor s koncovkou `.a` v nabízeném seznamu, a přidejte mu před `.a` ještě `_JP` (`J` - inicála jména, `P` - iniciála příjmení).
      - Po odsouhlasení jména na vás vyskočí okno `"Select resources to save"`. Klikněte na první řádek (soubor s příponou `.m`) - měl by se vysvítit. Pak zvolte tlačítko `"Change Filename"`, a rovněž u tohoto souboru připište k jeho jménu

"_JP" (těsně před příponu .m, obdobně jako u .a souboru). Po odsouhlasení zkontrolujte, že v okně "Select resources to save" je jméno správně upravené (s tím _JP). Odsouhlaste uložení tlačítkem "OK".

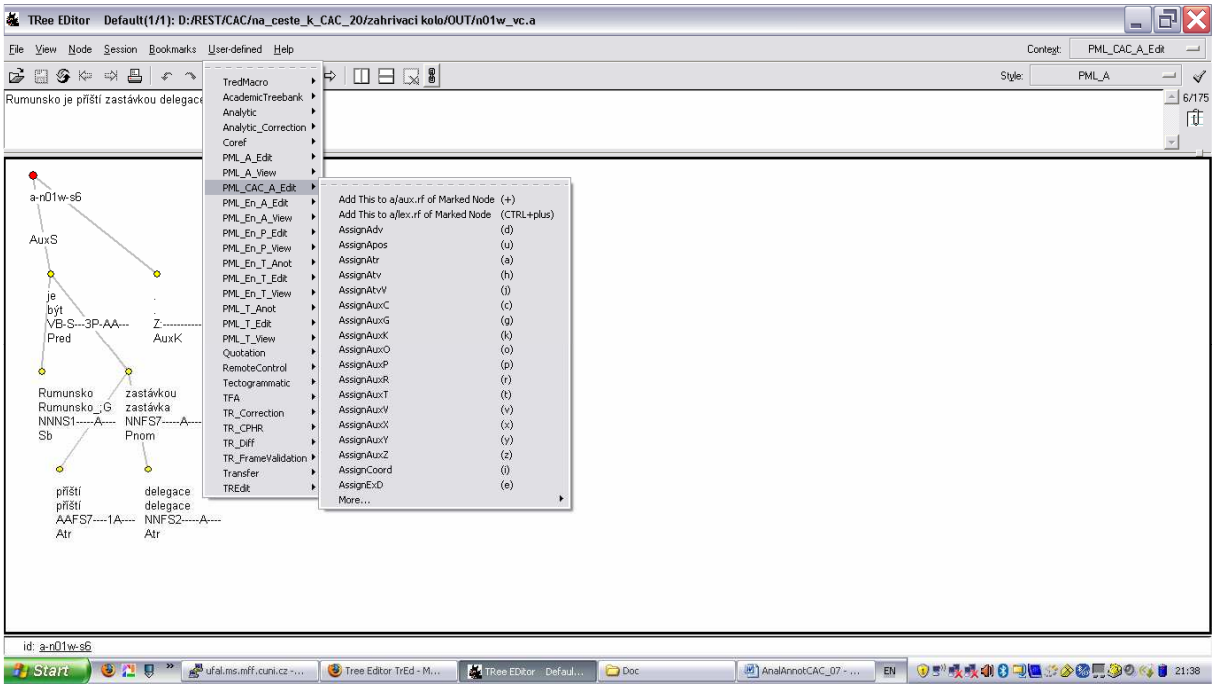
- Další otevření a uložení částečně anotovaného nebo dodatečně opravovaného souboru
  - Pro druhé a další otevření téhož souboru pro provádění další anotace nebo jakýchkoli jiných změn použijte soubor s modifikovaným jménem (...._JP.a).
  - Pro jeho uložení po provedení změn použijte File → Save (F2).
  - V okně "Select resources to save" klikněte na první řádek (mělo by v něm být jméno už rovněž modifikované, s _JP.m na konci). Hned poté odsouhlaste uložení pomocí tlačítka "OK" (tedy není nutno znovu měnit jméno souboru pomocí Change Filename).

- Zkušenosti říkají, že pro každý strom stačí, když u uzlu je zobrazeno slovo a jeho analytická funkce. K tomu, aby to tak bylo, je potřeba nastavit tzv. stylesheet , tj. to, co všechno se vám při anotaci zobrazí. Při všech dalších spuštění TrEd se bude zobrazovat to, co jste si nastavili. Z menu View->Edit Stylesheet , pravou část okna smažte a vložte následující, které potvrďte ok (viz Obrázek č. 3):

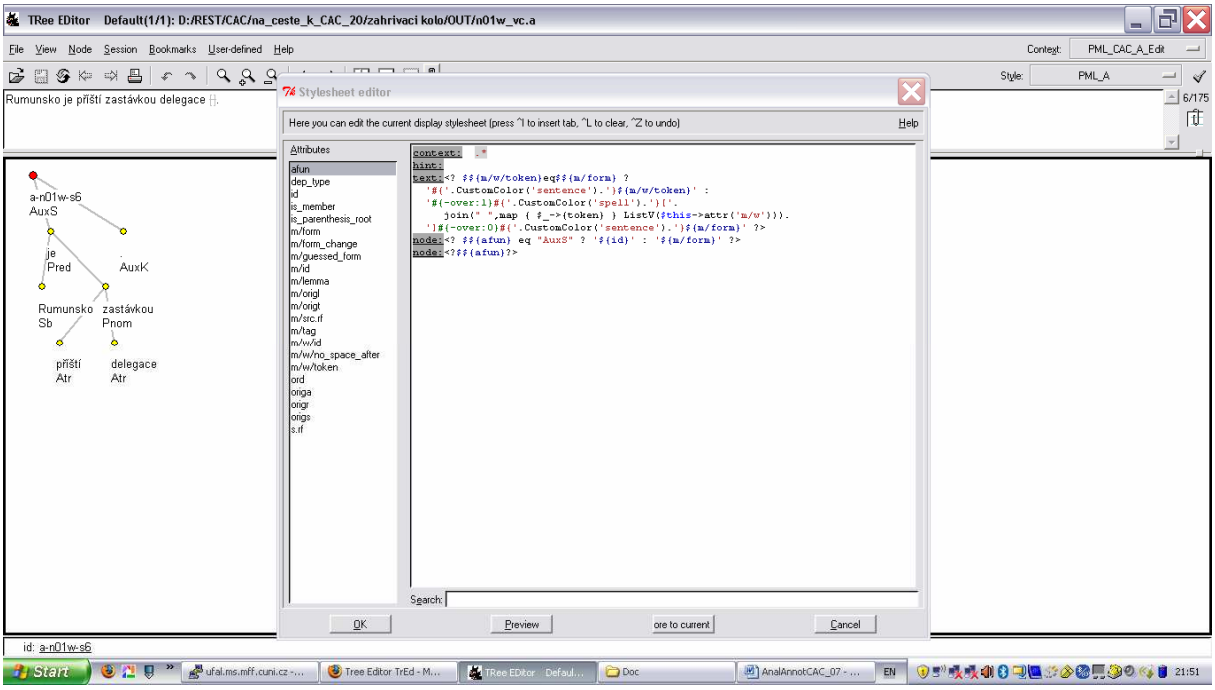
```
context: .*
hint:
text:<? $${m/w/token}eq$${m/form} ?
'#{'.CustomColor('sentence').'}${m/w/token}' :
'#{-over:1}#{'.CustomColor('spell').'}['.
join(" ",map { $_→{token} } ListV($this→attr('m/w'))).
']#{-over:0}#{'.CustomColor('sentence').'}${m/form}' ?>
node:<? $${afun} eq "AuxS" ? '${id}' : '${m/form}' ?>
node:<?$${afun}?>
```



Obrázek 1



Obrázek 2



Obrázek 3

6, Pokyny k anotaci

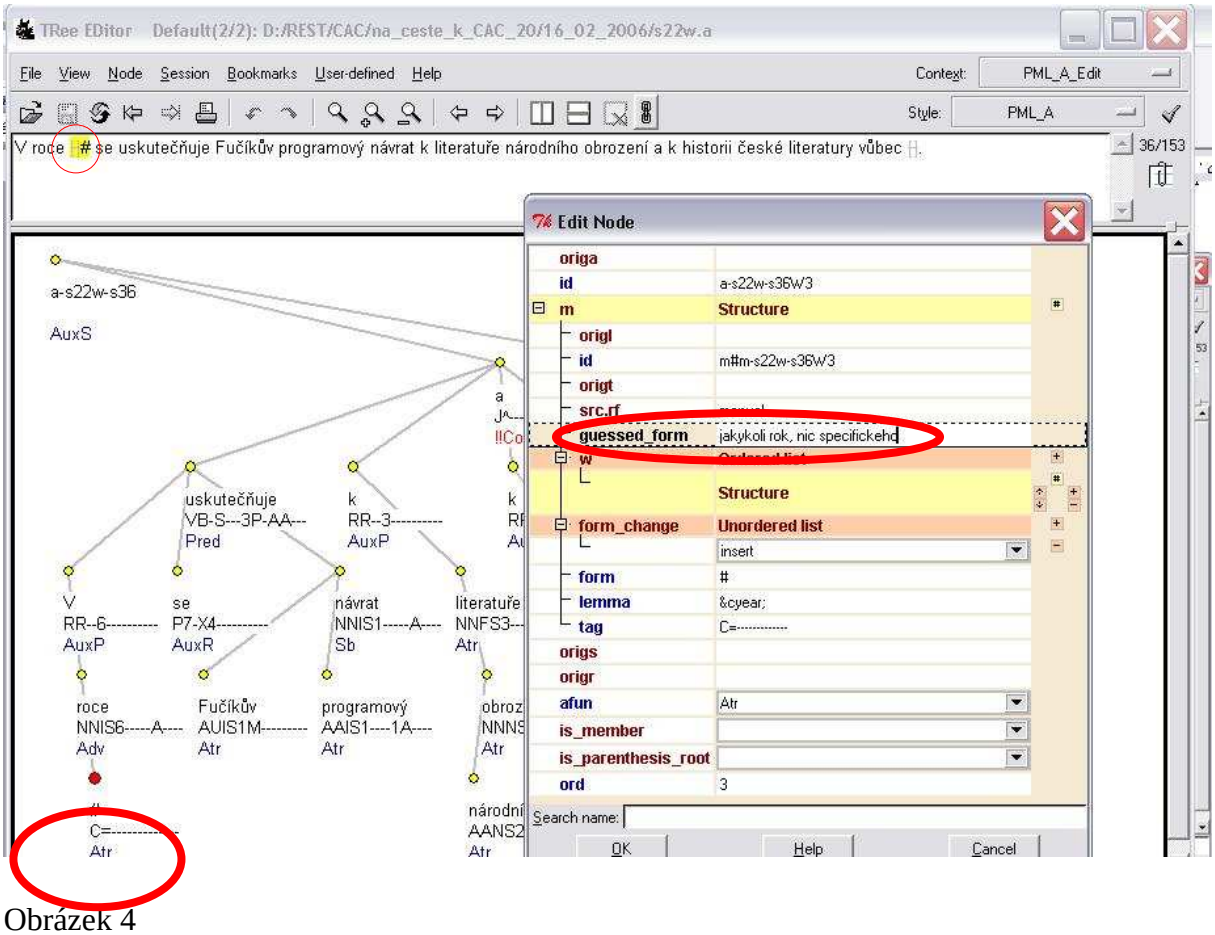
Anotace musí probíhat v souladu s anotačním manuálem (Hajič a kol., 2004). Samozřejmě, že manuál nepokrývá zvláštnosti ČAK. Proto upřesňujeme následující. Vybrané situace (viz seznam níže) NEOPRAVOVAT, pouze poznamenat název souboru, číslo věty a navrhované řešení. Zároveň oceníme jakýkoli názor na kvalitu stromečků a přiřazení analytických funkcí, které do anotace vstupují – tyto postřehy jsou důležité pro ladění automatické procedury parsingu. Poznámky přijímáme v elektronické podobě.

- chybně vložená interpunkce
- chybně specifikovaná cifra
- chybně specifikovaný typ cifry
- chybně specifikovaná podivnost

V některých kontextech je třeba si za jednotku, která byla do textu vložena (forma: #), případně označena jako podivnost (forma: ?), dosadit něco konkrétního, aby bylo možné větu syntakticky analyzovat. Informace o tom, co si anotátor sám pro sebe dosadil, je velmi důležitá. Proto je zaveden atribut `guessed_form`, který slouží k poznamenání právě toho, co si anotátor dosadil. Pro poznámky není definována žádná striktní syntax. Anotátor zadá buď přímo formu, pokud ji může uhádnout, nebo popis toho, co si myslí, že chybí, například "substantivum" (vzhledem k tomu, že nelze určit jaké substantivum tam bylo (třeba jméno)) – viz Obrázek č. 4.

NOVÉ: Atribut `guessed_form` používejte i v případě, když „něco“ není v pořádku s daným slovem, např. že čárka je tam, kde být nemá. Dále již toto nemusíte dokumentovat ve vašich osobních/poznámkových souborech.

Příklad:



Obrázek 4

7, Zahřívací kolo

1. **Prostudovat anotační manuál.**
   - html formát http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/html/index.html
   - pdf formát http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/pdf/a-man-cz.pdf)
2. **Shlédnout tutoriál k anotování**
   - http://ufallab.ms.mff.cuni.cz/video/recordshow/index/17/29
3. **Naučit se ovládat TrEd.**
   - nainstalovat TrEd dle pokynů ze strany http://ufal.mff.cuni.cz/~pajas/tred/
   - shlédnout tutoriál k TrEdu http://ufallab.ms.mff.cuni.cz/video/recordshow/index/2/23
4. **Seznámit se s valenčním slovníkem a návodem**, jak ho používat při syntaktické anotaci (viz část 8)

Pročíst anotační manuál celý najednou není asi zcela možné. Je dobré ho prolistovat, pak si pustit tutoriál, pak se k manuálu zase vrátit atd. Totéž platí i pro ovládání TrEd.

5. **Anotovat tři soubory.**
   Každý anotátor zpracuje v editoru TrEd nejvýše tři soubory, a to v tomto pořadí: `n01w` (175 vět), `s01w` (141 vět), `a03w` (123 vět). Tyto soubory byly zpracovány v rámci pilotní studie. Je tedy možné výstupy anotátorů porovnat s „pravdou“. Vstupní soubory budou předány elektronicky. Rovněž tak výstupní soubory spolu se soubory s poznámkami k anotaci. Název

výstupních souborů musí být sestaven dle pokynů uvedených v části 5. Soubory s poznámkami musí být pojmenovány dle tohoto klíče: soubor s poznámkami k anotaci souboru `n01w.a` anotátorem `XY` má název `xy_n01.*` (preferujeme pdf formát).

Jakmile anotátor zpracuje soubor n01w, tak patřičné soubory odešle na adresu hladka@ufal.mff.cuni.cz. My provedeme porovnání anotátorových anotací s „pravdou“, následně proběhne schůzka se zkušeným anotátorem – podrobnosti budou upřesněny. Pokud bude zkušený anotátor spokojen, anotátor začne anotovat „doopravdy“.

Balíček pro zahřívací kolo se jmenuje `first_CAK.zip`. Rozbalte ho do jednoho vámi zvoleného adresáře a spusťte TrEd, ve kterém si nastavíte, nebo už budete mít nastaveno vše, co je uvedeno v části 5, Otevřte v něm soubor `n01w.a` a můžete začít anotovat (kontrolovat)☺

8, Meetpoint

K projektu anotování Českého akademického korpusu máme založenou tzv. wiki stránku, na které je shromážděno vše, co k anotaci patří. K tomu, abyste se na stránku dostali, mě prosím kontaktujte (`hladka@ufal.mff.cuni.cz).`

Literatura

Český akademický korpus 1.0. http://ufal.mff.cuni.cz/rest/CAC

Hajič Jan, Jarmila Panelová, Eva Buráňová, Alevtina Bémová, Jan Štěpánek, Petr Pajac, Jiří Kárník. Anotace na analytické rovině: Návod pro anotátory. *TR-2004-23*, Ústav formální a aplikované lingvistiky, MFF UK, 2004.

Hladká Barbora, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, Jan Votrubech. *Průvodce Českým akademickým korpusem v. 1.0*. Nakladatelství Karolinum. 2007.

Hladká Barbora, Jiří Králík. Proměny Českého akademického korpusu. *Slovo a slovesnost 67*, s.179-194, 2006.

Pražský závislostní korpus 2.0. http://ufal.mff.cuni.cz/pdt2.0

Ribarov Kiril, Alla Bémová, Barbora Hladká. When the statistically oriented parser was more efficient than a linguist – a case on treebank conversion. *PBML 86*, 2006.

Příloha A Seznam souborů s počtem vět (jméno souboru:počet vět)

a01w:168	a02w:159	a03w:123	a04w:143	a05w:225
a06w:194	a07w:128	a08w:349	a09w:323	a10w:204
a11w:213	a12w:238	a13w:250	a14w:200	a15w:182
a16s:207	a17s:324	a18s:166	a19s:292	a20w:263
n01w:175	n02w:209	n03w:166	n04w:206	n05w:180
n06w:222	n07w:161	n08w:213	n09w:210	n10w:157
n11w:166	n12w:234	n13w:192	n14w:201	n15w:163
n16w:196	n17w:149	n18w:205	n19w:155	n20w:209
n21w:198	n22w:202	n23w:190	n24w:222	n25w:154
n26w:214	n27w:217	n28w:186	n29w:191	n30w:145
n31w:203	n32w:155	n33w:193	n34w:174	n35w:165
n36w:219	n37w:213	n38w:284	n39w:202	n40w:174
n41w:222	n42w:169	n43w:171	n44w:261	n45w:210
n46w:253	n47w:206	n48w:197	n49w:179	n50w:194
n51w:248	n52w:254	n53s:178	n54s:199	n55s:189
n56s:172	n57s:190	n58s:178	n59s:177	n60s:150
s00s:123	s01w:141	s02w:187	s03w:148	s04w:172

s05w:263	s06w:236	s07w:251	s08w:143	s09w:191
s10w:242	s11w:201	s12w:206	s13w:146	s14w:144
s15w:119	s16w:171	s17w:162	s18w:167	s19w:174
s20w:156	s21w:148	s22w:153	s23w:124	s24w:126
s25w:124	s26w:133	s27w:121	s28w:150	s29w:171
s30w:119	s31w:98	s32w:76	s33w:104	s34w:151
s35w:151	s36w:120	s37w:100	s38w:122	s39w:101
s40w:114	s41w:127	s42w:139	s43w:152	s44w:162
s45w:128	s46w:220	s47w:208	s48w:135	s49w:166
s50w:209	s51w:130	s52w:188	s53w:187	s54w:152
s55w:166	s56w:166	s57w:129	s58w:271	s59w:174
s60w:243	s61w:194	s62w:190	s63w:217	s64w:189
s65w:150	s66w:175	s67w:200	s68w:220	s69s:182
s70s:99	s71s:180	s72s:121	s73s:160	s74s:156
s75s:80	s76s:91	s77s:134	s78s:120	s79s:161
s80s:147	s81s:171	s82s:143	s83s:163	s84s:231
s85s:101	s86s:107	s87s:179	s88s:105	s89s:171
s90s:102	s91s:144	s92s:143	s93s:199	s94s:152
s95s:106	s96s:170	s97s:132	s98s:129	s99s:174