ANOTÁTORSKÁ SCHŮZKA 17. 5. 2023

Anotátorský tým:
Iva Hrabánková
Karolína Churáčková
Lenka Charvátová
Leona Špetová
Tereza Marková
Tadeáš Tajbl
Tomáš Harcuba

Stav anotace
Data PCEDT jsou hotová.
Data PDTSC jsou hotová.
Anotujeme korpus **PDT**: články z Lidových novin, časopisu Vesmír atp. Hezká data.

~~PCEDT: 49 208 vět~~
~~PDTSC: 73 835 vět~~
FAUST: 3 000 vět (chybí anotace koreference)
PDT tamw 49428 vět
amw 38479
mw 27710
Stav 20. 4.: Zbývá ještě 2631 přiřazených a 43 728 nepřiřazených vět z tamw.

Srovnání korpusů. Průměrná délka věty.

corp	avg (počet tokenů dělený počtem vět)
PCEDT	23.40
PDTSC	10.05
PDT	16.86

Sazba za větu
Z dat od anotátorů plyne **rychlost anotace PDT**:
1,4 věty/min -> 85 vět/hod -> 1,8 Kč/věta (150 Kč/hod)
Pravidelná anotace 2-3 hodiny týdně - > 2 000 vět /měsíc
Sazba od prosince 2022 (zůstává):
do 1500 vět za měsíc: 1,50 Kč/věta
1500-2500 vět za měsíc: 2 Kč/věta
přes 2500 vět za měsíc: 4 Kč/věta

Data PDT
A-rovina již anotována ručně podle původního manuálu.
- Kontrola, zda anotace je v souladu s t-rovinou
- Zohlednění nových pravidel. Používat automatické kontroly.

Na t-rovině je více anotace než v předcházejících korpusech.
- lemmatizace
  -- *kdo, někdo, nikdo, kdosi* -> kdo
  -- *tři, trojí, třetina, třetí* -> tři
  -- *pěkně* -> pěkný
- valence substantiv:
  --*výroba nábytku*.PAT vs. *výroba nábytku*.RSTR
  -- *košík hub*.MAT vs. *košík hub*.RSTR
  -- *otec Pavla*.APP vs. *otec Pavla*.RSTR