Aktuální úkoly v získávání dat, Corpus Architect

Vít Suchomel

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

4. prosince 2012

INVESTMENTS IN EDUCATION DEVELOPMENT

Osnova

- Corpus Architect – konference Gramatika a korpus
- Corpus Architect – extrakce termů
- seznamy slov z matematických dokumentů
- czTenTen12

Corpus Architect – konference Gramatika a korpus

- prezentace 30. 11.
- narozdíl od ostatních není výzkum
- narozdíl od ostatních mnoho dotazů
- zájem UPOL (korpusy, czTenTen12, zabudované značkování)
- zájem o rozhraní k paralelním korpusům

Corpus Architect – Ajka + Desamb

- již brzy :-)
- testovací verze na `http://ske.fi.muni.cz`
- chceme volitelný převod do pražské notace značek?

Corpus Architect – extrakce termů

- dvojice slov ve skečových relacích
- základ naprogramoval Honza P., dokončil jsem já
- commonest match – Vítek B.
- ukázky

Seznamy slov z matematických dokumentů

- pro EUDML k lepšímu rozpoznávání znaků
- počáteční slova → WebBootCaT → wordlist
- počáteční slova z 2010 Mathematics Subject Classification ... 41 množin slov
- cca. 800 dotazů do Bingu ... 5093 dokumenty, 31.5 M tokenů

Jak získat počáteční slova v jiných jazycích?

- extrakce klíčových slov ze stávajících dokumentů v EUDML
- překlad MSC

czTenTen12

- včetně skečů
- slovo "rozhlas" je lemmatizováno jako "rozhlásit" jen v 5 % případů
- "přede vším" již není lemmatizováno "přís̈t vš̈i"
- zkoušejte na `http://ske.fi.muni.cz`