Ukázka nového rozhraní paralelního korpusu

Ondřej Mrázek

Filozofická fakulta
Ústav českého jazyka

Masarykova univerzita, Brno

Obecný popis

- rozhraní zcela přepracováno (Backend i Frontend)
- navržena nová struktura pro hierarchizaci textů
- nové texty (Bible Kralická, Melantrichova)
  - korekce přepisů
  - stanovována pravidla pro přepisy
- formální popis struktury dat XML (příprava k dokumentaci)

Použité technologie

- serverové fce pro práci s textem v jazyce python (dříve bash skript)
- rozhraní bootstrap (dříve čisté html a css)
- js (jQuery + AJAX) (dříve js, cgi skripty)
- výhody:
  - standardní způsob
  - lepší čitelnost a správa kódu
  - snazší úprava rozhraní (efektivnější)

Ukázka práce v rozhraní

- procházení po verších
- paralelní porovnání
- možnosti nastavení
- vyhledávání:
  - typy dotazů
  - ukázka vyhledávání (regexy)
  - paralelní porovnání
  - zobrazení konkordance a počtů

Hyperword

- odlišný přístup než hyperlemma
- hyperlemma: musí vygenerovat sl. tvary včetně hláskových alter. během vývoje (řešeno jinde, otázka úspěšnosti)
- hyperword: dotaz word rozšířený pomocí gramatiky (kůň: kůň, kuoň; ou – au, v – w, ů - ou -ú, přidání příklonného ť, ž aj.)
- pravidla náhrady a podmínek pozic náhrady (není možné nahradit ve všech pozicích)
- návrh základní gramatiky, rozšířené gramatiky, míra úspěšnosti

Značkování

- idea: Je možné, že i na texty ze starších fází češtiny lze použít novočeský morfologický analyzátor s minimem úprav dat při zachování obstojné úspěšnosti.
  - ani novočeský text není stoprocentně označkovaný
  - frekvence problematických jevů se přímo úměrně zvyšuje se stářím textů
- majka, desamb
- problém u starších textů:
  - zjistit míru neúspěšnosti dle stáří
  - najít problematické jevy, navrhnout úpravu dat

1. preprocessing: použití gramatiky na "první"sloupec vertikály (lhauce – lhouce)
2. značkování
3. postprocessing: přidání pravidel pro již desambiguovaný text (druhá desambiguace), náhrada některých informací (pravidlový přístup)