CJBB75 - Základy práce s korpusem

3.10.2012 - 1. hod

# Korpus

- elektronický

- representativní (celý jazyk, úsek jazyka)

- anotace (typ textu, charakteristiky, doba vzniku, ..., vnitrotextové anotace - popsány jednotky textu)

- tokenisován - segmentace

Korpusové manažery: např. Bonito, webová rozhrání

## Korpusy synchronní

- k disposici 3 - např. syn2005 (po 10 letech)

- stejný rozsah - stomilionů slov

- nemají stejné složení, nelze například dělat stats ze slovní zásoby (syn2000 má většinu z žurnalistiky, některý jiný třeba z beletrie)

- dva korpusy publicistické - syn200x-pub

## Anotace

- vnější anotace, morfologické značkování

- každá jednotka analysovaná jakožto slovní tvar/interpunkční znak má přiřazenou charakteristiku lemma (zákl. slovníkový tvar) a značka (slovníková interpret. konkr. slovního tvaru)

- v korpusu lze vyhledávat podle všech tří parametru (anotace, morf. značk., ??)

**Lemmatisace** - pokud je korpus lemmatisovaný, lze vyhledávat v základním tvaru (*pes* vrátí *psy, psů, pes*, ...)

## Mluvené korpusy

pražský mluvený korpus, brněnský ~

- pražský už před rokem 1990

oral2006, oral2008 - mapují různé oblasti Čech (a část. Moravy)

- rozsah pouze v milionech

**ukol: (v tydnu, kdy neni hodina)**

navrhnout postup, jak vyhledat v korpusu syn2010 tvary masculin v genitivu pluralu (pomoci /bez pomoci tagu - morf. zn.)