KORPUSNO JEZIKOSLOVJE

Uvod

Tehnične informacije:

- 30 ur predavanj, 30 ur vaj
- 6 ECTS
- dva kolokvija ali pisni izpit ali seminarska (podiplomci)
- domače naloge (20 % skupne ocene)
- obvezna prisotnost pri vajah (80 %) - pogoj za pristop k pisnemu izpitu/oddajo seminarske

Vsebina predmeta:

Predavanja:

1. Uvod – pregled predmeta: informacije o predmetu, kaj so korpusi
2. Raba korpusov: seznami, statistične metode
3. Gradnja korpusov
4. Označevanje korpusov:
5. oblikoslovje, skladnja, ...
6. Zapis znakov:
7. Unicode
8. XML

Vaje:

1. Pregled obstoječih korpusov; Fidaplus podrobno
2. Raba korpusov: WordSmith 2x
3. Raba korpusov: Wordsketchengine
4. Izdelava korpusa: Obuti maček 2x
5. Delo s pridobljenim korpusom

Kaj je korpus

- obsežna zbirka besedil
- jezik v resnični in sodobni podobi
- elektronska oblika
- vzorec jezika, ki naj bi ga predstavljal
- služi za opisovanje jezika

Vrste korpusov:
- Medij: pisana in govorjena besedila
- Obseg: referenčni korpusi, korpusi podjezikov
- Časovni razpon: diahroni in sinhroni
- Jezik: enojezični in večjezični:
  - vzporedni
  - primerljivi

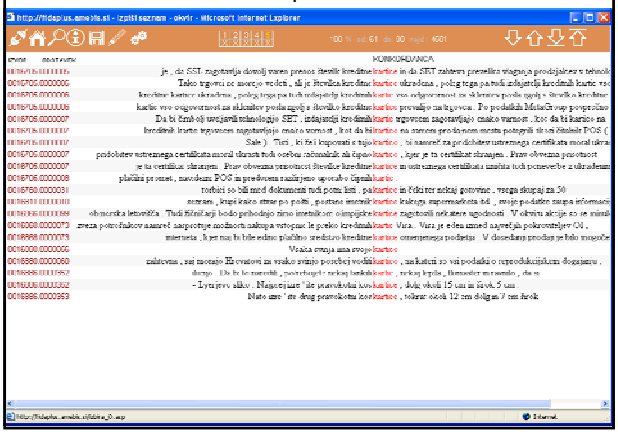
Zakaj potrebujemo korpuse?

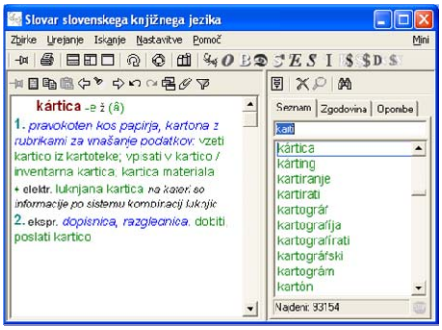
- Izdelava slovarjev in drugih jezikovnih virov
- Izdelava slovnic in drugih opisov jezikovne strukture
- Razvoj pripomočkov za prevajanje
- Izdelava pripomočkovza učenje jezika
- Jezikovne tehnologije
- Raziskovanje vseh oblik jezikovnega vedenja

Kako uporabljamo korpuse?

- **besedni seznami:**
  Katere besede so v korpusu? V posameznem besedilu? Katere izstopajo po pogostosti uporabe?
- **konkordance (**opazovanje besed skupaj s sobesedilom): kako so besede uporabljene? kaj torej pomenijo?
- **statistične metode:**
  opazovanje zanimivih sopojavitev besed (kolokacije), narativne študije, ...

"kartica" v korpusu FidaPLUS





Zakaj nam tega ne pove slovar?

Orodja za analizo korpusov

- Veliki korpusi so dostikrat na medmrežju, skupaj s svojimi vmesniki: BNC, FidaPLUS
- Verjetno najboljši medmrežni vmesnik: SketchEngine
- Kupljeni programi na lastnem računalniku
  npr. WordSmith (in seveda korpus!)
- Izdelava lastnih programov:
  npr. Perl, R
- Izdelava lastnih korpusov:
  ročno, BootCat

Gradnja

Če ustreznega korpusa ni na voljo, ga moramo narediti sami

Postopek:
1. izbira besedil: reprezentativnost, uravnoteženost, izvedljivost
2. digitalni zajem: OCR, Word, HTML
3. normalizacija besedil: enovit format
4. (označevanje: oblikoslovne oznake, lematizacija)
5. (distribucija: avtorske pravice, platforma)

Označevanje

Korpus je lahko precej bolj uporaben, če je jezikoslovno označen

Ravni označevanja:
- oblikoskladenjske oznake (samostalnik, ženski spol, ednina, rodilnik)
- leme
- skladenjsko označevanje (stavčni členi)
- besedilno označevanje (anafore ...)
- ...

Unicode

Kako so v besedilih kodirani znaki?
Zakaj je to zanimivo?
- kadar gre kaj narobe in č postane c, ali pa kaj drugega
- kadar je potrebno uporabljati nenavadne znake (npr. dolgi s, fonetično abecedo, ...)

Obstaja veliko starejših kodnih naborov (ki pa se še uporabljajo), moderna tehnologija pa uporablja univerzalen (pa za razmeroma kompleksen) nabor znakov unikod (Unicode)

XML

Kako naj bi bilo korpusi zapisani na računalniku, da bodo uporabni za uporabnike z različno računalniško/programsko opremo in za različne namene?

Standardizacija na spletu:
- izmenljivost, trajnost, uporabnost, razumljivost, preverljivost
- konzorcij W3C
- standardi konzorcija W3C (SGML, HTML, XML)
- eXtended Markup Language (podmnožica SGML)

Posebej za zapis besedilnih korpusov:
- TEI (Text Encoding Initiative), od l. 2002 zapisan v XML