Vaje VI

Izdelava korpusov:
izbira besedil, označevanje, instalacija na SKE

Gradnja korpusa po korakih

1. Zbiranje besedil v različnih formatih, preoblikovanje v enotni format (txt) + enotni kodni zapis (najbolje UTF-8):
   - lahko naredimo sami
   - ali pa za nas naredi BootCat
2. Označevanje:
   tokenizacija in segmentacija, oblikoskladenjsko označevanje, lematizacija
   - spletni servis
     http://nl.ijs.si/jos/analyse/
3. instalacija korpusa v konkordančnik
   - Sketch Engine

WebBootCat

1. Izberemo ključne besede
   (več ko je ključnih besed, večji bo korpus in dalj bo trajala gradnja)
2. Nastavimo jezik na slovenski, mogoče spremenimo prednastavljene parametre (npr. število strani po poizvedbi, če hočemo večji korpus, vendar potem gradnja traja dlje)
3. Pregledamo najdene domače strani
   (ali pa vzamemo vse)
4. BootCat izdela korpus
   (za slovenski jezik neoznačen)
5. Ko je korpus izdelan, nas BootCat o tem obvesti po emailu
6. Ta korpus lahko neoznačen že kar uporabljamo
7. Če pa hočemo korpus jezikoslovno označit:
   - najprej shranimo korpus na našem računalniku v "*raw format*"

Označevanje

- Korpus označimo preko spletnega servisa http://nl.ijs.si/jos/analyse/
- Oblikoslovne oznake so po specifikaciji JOS
- Podobne, vendar razne sprembe glede na oznake FidaPLUS!
- Podrobnejši pregled: http://nl.ijs.si/jos/msd/html-sl/

Instalacija korpusa

Označeni korpus naložimo na SketchEngine:

- Izberemo CorpusBuilder, "Create new corpus: from template"
- Nastavimo opcije:
  - Tagged WS
  - (Uploaded files metadata: Title)
  - Uploaded files encoding: UTF-8
- Korpus spustimo skozi vse korake instalacije (merge, vert, …)
- Naknadno lahko dodajamo nove podkorpuse našemu korpusu

Uporaba

- Uporabimo na novo instaliran korpus
  - konkordance
  - word-sketches (pri majhnih korpusih zmanjšamo spodnjo število najdenih primerov iz 5 na npr. 3)
  - tezaver
  - sketch differences

Problemi s avtomatskim označevanjem

- Problemi z razdvoumljanjem:
  - Jesti vs. biti; elativ (preostalo)
- Problemi z neznanimi besedami:
  - Memo, lematiziran kot "meti", tajkun,
- Problemi predvsem tam, kjer se tudi pri ročnem označevanju ne znamo prav dobro odločiti
  - eni/prvi – drugi, pridevniki vs. deležniki na –n, …

Fida+: iskanje po 2. kanalu

Primeri:

- Glagol brati – s katerimi predlogi se veže?
- Najdite vse tožilniške predložne besedne zveze, ki sledijo glagolom "-pisati" (glagol+predlog+samostalnik v tožilniku). Prejšnji iskalni niz nadgradite z omejitvijo, da naj bodo glagoli v velelniku. Zdaj pa prejšnjemu iskalnemu nizu dodajte še pogoj, da predlog ne sme biti "na".
- Sam. im. sr. sp. mn + sam. im. ž. sp. mn. + gl. del. mn. – kateri spol?
- Kakšno obliko ima množilni številnik, izpeljan iz števila 7, v dvojini? In kakšno številnik, izpeljan iz števila 8, 2, 3? Kako pogosta je raba množilnega številnika? Kateri številniki so uvrščeni med 'druge'?
- Kako bi našli pojavitve glagolskega časa predpreteklik v korpusu?

SKE: iskanje s pomočjo oznak

- Osnove CQP skladnje (samo toliko, da boste znali poiskat v SKE tudi po oznakah):

http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPSyntax.html

http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPExamples.html

- Primer: [lemma="zadnji.*" & tag="So.*"]