Vaje V

Sketchengine

Raba

Izdelava korpusov (izbira besedil, označevanje, instalacija na SKE)

SKE

- Konkordanca: primerjava med stranjo od Fide+ in SKE: kaj se kje da/ne da narediti:
  - Pri Fidiplus lahko poiščemo več različnih besed levo in/ali desno od iskalnega pogoja, pri SKE samo eno.
  - SKE istočasno in mnogo hitreje
- Primerjava med BNC in F+ (različne možnosti iskanja)
- Shranjevanje podatkov
- Grajenje lastnih korpusov – naslednjič.

- Word sketches:
  - mrč,veleum, tajkun,
  - jesti, driblati
- Thesaurus: poišče besede, ki imajo podobno distribucijo in vezljivost kot iskana beseda
- Sketch diff: pokaže razlike in podobnosti med dvema besedama
  - cona/območje/področje
  - baba/mačka/bejba
  - dojenček/dojenec/malček/otrok

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://nl2.ijs.si/analyze/
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://nl2.ijs.si/analyze/
- Oblikoslovne oznake so po specifikaciji JOS
- Podobne, vendar razne spremembe 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, …

SKE: iskanje s pomočjo oznak

- Osnove CQP skladnje:

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.*"]