Korpusno jezikoslovje
2009/2010

Univerza v Novi Gorici
Fakulteta za humanistiko

1. kolokvij 29/03/29, 15:00 - 16:30

REŠITVE

1. naloga (2 točki)

Opišite EVROKORPUS, http://evrokorpus.gov.si/ pri čemer upoštevajte tipologijo in značilnosti korpusov.

Odgovor:
EVROKORPUS je zbirka vzporednih dvojezičnih korpusov prevodov, ki je bila narejena iz pomnilnikov prevodov, ki so nastali v Sektorju za prevajanje Službe Vlade Republike Slovenije za evropske zadeve. Vsebuje angleško-slovenski, nemško-slovenski francosko-slovenski, italijansko-slovenski in špansko-slovenski (pod)korpus, vsega skupaj okoli 130 milijonov besed. Korpus je namenjen zlasti prevajalcem strokovnih besedil in je dostopen preko spletnega vmesnika.

EVROKORPUS je torej večjezični vzporedni pisni korpus in korpus podjezika, saj vsebuje samo besedila, ki so povezana s pravnim redom EU. Ker je nastal na osnovi pomnilnika prevodov, ga sicer lahko imamo za vzorčni korpus, saj pomnilniki prevodov tipično ne vsebujejo celotnih besedil; vendar pa vsebujejo večino besedil, zato se kot pravilen odgovor šteje tudi, da je celovit korpus. Korpus se občasno posodobi, zato bi ga pogojno lahko imenovali spremljevalni korpus. Korpus je neoznačen, saj besedila v njem nimajo jezikoslovnih oznak.

2. naloga (3 točke)

Opišite, kako bi v korpusu FidaPLUS našli primere rabe predpreteklika. Razmislite, kakšna je formalna skladenjska oblika predpreteklika, in določite, kakšen bi bil osnovni iskalni pogoj ter kaj bi poiskali v situ.

Odgovor:
Fomalna oblika predpreteklika: vezni glagol v sedanjiku+vezni glagol v pretekliku (ujemanje v spolu in št.) + deležnik –l. En primer uporabe predpreteklika bo napisan na tablo.
Osnovni iskalni pogoj: poiščeš deležnike (#2Gpd... – točnega zaporedja oznak na pamet ne bi zahtevala, ker se mi zdi učenje oznak na pamet brezvezno).
Sito: poiščeš skupke veznega glagola v sedanjuku + pretekliku eno ali dve besedi levo od deležnika (#2Gvs..._#2Gp...)

3. naloga (2 točki)

V korpusu http://nl2.ijs.si/dsi.html poiščite zaporedja štirih zaimkov, pri čemer naj bo prvi zaimek dolg natančno dve črki, drugi pa tri. Koliko takšnih lem (pojavnic in različnic) vsebuje korpus?

Odgovor:
V vrstici »bes.vrsta« izberemo zaimek za prve štiri pojavnice. V vrstici »beseda« pojavnici 1 vpišemo »..«, pojavnici 2 pa »…«.Ker nas zanima število pojavnic in različnic za »Prikaz« izberemo »Seznam«.
Ker nas zanimajo leme, v stolpcu »Prikaži« odkljukamo »lema«. (tudi brez tega pogoja se je odgovor štel kot pravilen).
Odgovor je 7 različnic in 17 pojavnic.

Query: IKORPUS; [word="." & pos="zaimek"] [word="..." & pos="zaimek"] [pos="zaimek"] [pos="zaimek"]

N°	Hits	Atts	Hit
1	5	**lemma**	se ti ves ta
2	4	**lemma**	se ti kaj tak
3	2	**lemma**	se kdo kaj tak
4	2	**lemma**	se jaz ves ta
5	2	**lemma**	se jaz kaj takšen
6	1	**lemma**	ta kar se on
7	1	**lemma**	se ves kar se

7 types, 17 tokens

4. naloga (3 točke)

Opišite, kako bi na najhitrejši način poiskali možne kandidate ključne besede v nekem besedilu.

a. Kateri program bi uporabili v ta namen?
b. Kakšen je postopek za iskanje ključnih besed v tem programu?
c. Kakšne so slabosti in pomanjkljivosti iskanja s tem programom?
d. Kaj je to stoplista in ali je pri iskanju ključnih besed uporabna?

Odgovor:
Wordsmith. Primerjava seznama besed v obravnavanem korpusu s seznamom besed v referenčnem korpusu. Ključne besede poišče zgolj s pomočjo primerjanja relativnih pogostosti pojavljanja besed v korpusu; ključne besede so lahko tudi relativno nepogoste besede. Stoplista je seznam najbolj pogostih, običajno funkcijskih besed v jeziku, ki nas pri izdelavi besednih seznamov in seznamov ključnih besed običajno ne zanimajo. Če jo vklopimo pri iskanju ključnih besed, pa je lahko tudi relativno neuporabna, ker besede, ki so absolutno najbolj pogoste v obeh korpusih, izloči tudi že sama formula, po kateri so ključne besede izračunane.