Korpusno jezikoslovje
2008/2009

Univerza v Novi Gorici
Fakulteta za humanistiko

1. kolokvij 28/11/08, 11:30 - 12:15

1. naloga (2 točki)

Opišite korpus ICE, http://www.ucl.ac.uk/english-usage/ice/, pri čemer upoštevajte tipologijo in značilnosti korpusov.

Odgovor:
ICE je korpus lokalnih različic angleškega jezika, ki ga trenutno pripravlja 18 raziskovalnih skupin po svetu. Vsak podkorpus ICE ima milijon besed, in je vzorčen, ker vsebuje iztržke iz besedil, podkorpusi pa so med seboj primerljivi. Je sinhron korpus, in vsebuje besedila od 1989 naprej. Je tako pisni kot govorjen korpus. Korpus ni reprezentativen za posamezna govorna področja, vsebuje pa širok razpon besedilnih tipov, zato bi ga lahko smatrali za referenčnega. Korpus je označen oblikoslovno in skladenjsko in je prosto dostopen.

2. naloga (3 točke)

V korpusu FidaPLUS najdite primere rabe deležnika na –vši. Ker je najpogostejša beseda na –vši beseda 'bivši', jo v iskalnem pogoju izločite (notranje zanikanje). Med vsemi najdenimi zadetki izberite tiste, ki se nadaljujejo s samostalnikom v tožilniku (#2So??t*). Kako pogosto se take zveze uporabljajo?

Odgovor:
Iskalni pogoj je *vši&~bivši (lahko pa tudi +vši&~#2p). Rezultatov je 531. Ko nam najde vse besede, ki ustrezajo iskalnemu pogoju, v situ določimo, da nas zanimajo rezultati, pri katerih je prva naslednja beseda (1. beseda od 01 do 01) #2So??t*. Takih rezultatov je 60, od tega 53 zvez z besedo *vštevši*.

3. naloga (3 točke)

V korpusu http://nl2.ijs.si/dsi.html poiščite besede, ki imajo tri soglasnike ali več a kupu ("težki zlogi"), vendar ne smejo stati na začetku besede. Napišite regularne izraze, ki ste jih uporabili in koliko pojavnic in različnic ste našli v korpusu.

Odgovor:
Iskalni izraz:

.+[bcčdfghjklmnprsštvzž]{3,}.*

oz., če bi bilo zelo natančni in bi hoteli zajeti tudi besede, ki so iz samih velikih črk:

.+[bcčdfghjklmnprsštvzž*BCČDEFGHJKLMNPRSŠTVZŽWQXY*]{3,}.*

Oba izraza vrneta toliko zadetkov, da dosežemo omejitev konkordančnika na 100,000 pojavnic. Od teh je različnic 16630.
Iskalni izraz lahko tudi skrajšamo z uporabo negacije:

.+[^aeiouAEIOU1-9]{3,}.*

4. naloga (2 točki)

Opišite, kako bi na najhitrejši način poiskali ključne besede v nekem besedilu. Kateri program bi uporabili v ta namen? Kakšen je postopek za iskanje ključnih besed v tem programu?

Odgovor:

Če nas zanima ključnost v smislu pogostosti pojavljanja neke beseda ali besedne zveze v nekem besedilu v primerjavi z drugimi besedili, potem lahko uporabimo avtomatsko iskanje ključnih besed, kakršnega nudi orodje Keywords v Wordsmith tools. Ključne besede poiščemo tako, da manjši, bolj specializiran korpus, primerjamo z večjim korpusom, ki nam v tem primeru služi kot referenčni (v ta namen uporabimo besedna seznama, ki smo ju izdelali iz teh dveh korpusov). Besede, ki se v specializiranem korpusu pojavljajo relativno bolj pogosto kot v referenčnem, so v programu predlagane kot ključne besede.