1.2 Korpusi in korpusno jezikoslovje – vaje

Besedilni korpusi so obsežne zbirke realnih besedil v elektronski obliki. Besedila so zajeta iz različnih virov na način, da predstavljajo **vzorec jezikovne rabe** določene vrste. Korpusna besedila tipično vsebujejo programsko ali ročno pripisane **oznake,** npr. osnovno obliko besede, besedno vrsto in druge lastnosti besede. Za raziskovanje besedilnih korpusov so besedila umeščena v **konkordančnike:** specializirane programe, ki omogočajo napredno iskanje po besedilih, razvrščanje, filtriranje, izvažanje podatkov in podobno.

Da lahko pravilno interpretiramo in generaliziramo ugotovitve, moramo dobro razumeti, **kakšna besedila** določen korpus vsebuje, kako je bil **zgrajen** in kakšen je njegov **namen.**

Besedilne korpuse uporabljamo:

- ker naša **jezikovna intuicija** ne more natančno predvideti, kako se jezik v širši rabi obnaša,
- ker s pomočjo računalnika lahko obdelamo večje količine podatkov na naprednejše načine in tako lažje poiščemo relevantne **jezikovne vzorce in trende**,
- ker so zgrajeni na transparenten in dokumentiran način, da lahko podatke ustrezno **interpretiramo** in **generaliziramo.**

Korpusi se uporabljajo za različne namene v **uporabnem jezikoslovju** (za pripravo slovarjev, slovnic, šolskih gradiv ipd.), **teoretičnem jezikoslovju** (za raziskave, ki lahko vodijo do novih dognanj o jezikovni rabi in sistemu), pri drugih poklicih, ki se posvečajo **pisni produkciji** (pisanje, prevajanje, lektoriranje ipd.) in tudi za **ljubiteljsko raziskovanje** jezika (preverjanje jezikovne rabe, raziskovanje raznih zanimivosti ipd.)

Za slovenščino trenutno še ne obstaja veliko priročnikov, ki so narejeni na osnovi korpusnih podatkov (v prihodnosti jih bo več). Korpusi so tudi sodobnejši od nekaterih obstoječih priročnikov, zato se korpusni podatki in podatki v priročnikih mestoma razlikujejo). V praksi se korpusi zato pogosto uporabljajo kot dopolnilo obstoječim jezikovnim priročnikom.

Za slovenščino je na voljo več različnih korpusov. Na taboru bomo natančneje spoznali naslednje:

IME KORPUSA	VRSTA JEZIKA, POVEZAVA	ZAJETA BESEDILA
Kres	Splošna pisna slovenščina	časopisi, revije, leposlovje, strokovna literatura, spletna besedila, besedilni drobiž
GOS	Govorjena slovenščina	televizijske in radijske oddaje, javni nastopi, sestanki, zasebna komunikacija …

Janes	Spletna slovenščina	tviti, blogi, uporabniški komentarji, forumi
Šolar	Jezik šolarjev	šolski eseji in testi + učiteljski popravki

Gigafida je obsežna zbirka sodobnih (1990-2011) slovenskih besedil iz časopisov, revij, knjig, s spleta itd. Korpus obsega skoraj 1,2 milijarde besed. **Kres** je manjša različica tega korpusa, prinaša cca. 100 milijonov besed. Korpuse, ki prinašajo splošni jezik, imenujemo **referenčni korpusi.** Ti se uporabljajo za izdelavo referenčnih priročnikov, v raziskavah pa jih pogosto uporabljamo tako, da z njimi primerjamo rezultate iz drugih korpusov.

GOS je prvi korpus govorjene slovenščine. Prinaša posnetke govora v različnih vsakodnevnih situacijah. Posnetki so **transkribirani** in umeščeni v zmogljiv konkordančnik, s katerim lahko primere govora iščemo, poslušamo in preučujemo. Korpus obsega okrog **milijon besed.** Namenjen je raziskovanju govora.

Šolar vsebuje pisna besedila, ki so jih učenci in dijaki slovenskih šol tvorili pri pouku. V precejšnjem delu besedil so posebej označene tudi jezikovne napake, ki so jih v spisih **popravili učitelji.** Po slednjih lahko s pomočjo specializiranega konkordančnika tudi iščemo. Korpus vsebuje približno **milijon besed,** namenjen je raziskavam šolske pisne produkcije oz. jezikovne zmožnosti šolarjev in pripravi učnih gradiv.

Janes je korpus spletne slovenščine. Vsebuje besedila, ki so jih na spletu tvorili uporabniki, in sicer tvite, forumska sporočila, blogovske zapise in komentarje spletnih novic. Korpus obsega okrog **134 milijonov** besed. Namenjen je raziskovanju nestandardne spletne slovenščine. Korpus je eden od rezultatov nacionalnega raziskovalnega projekta *Jezikoslovna analiza nestandardne slovenščine* (J6—6842), ki poteka med leti 2014 in 2017, v njegovem sklopu pa je organiziran tudi naš poletni tabor.

1.2.1 Od konkordance do kolokacije – prvi del

1.2.1.1 Korpus KRES

http://www.korpus-kres.net/

1. Odpremo korpus Kres in vtipkamo v iskalno okence besedo *pljuvalnik.* Ogledamo si rezultate v konkordančniku in spoznamo:

- kaj je **konkordanca** oz. **konkordančni niz, konkordančno jedro,**
- kje najdemo število konkordanc,
- kako pridemo do širšega **sobesedila, metapodatkov o** besedilu in **korpusnih oznak,**
- kaj so **filtri** in kako jih uporabljamo.