Špela Arhar Holdt (spela.arhar@trojina.si) in Jaka Čibej (jaka.cibej@gmail.com)

Korpusi in korpusno jezikoslovje

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.

Od konkordance do kolokacije – prvi del

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.

2. Raziščemo, kaj pomeni beseda *pljuvalnik.* Kaj pomeni *brbotalnik?* Poznamo sopomenko za brbotalnik? Kako pogosto in v katerih besedilih se pojavlja *brbotalnik,* kako pogosto pa sopomenka?

3. Katere besede na *–nik* še poznamo? Naštejemo tiste, za katere mislimo, da so v korpusu najpogostejše. Odpremo zavihek Seznam in vnesemo iskalni pogoj **nik.* Ogledamo si rezultate. Ogledamo si filter Besedna vrsta in komentiramo, kar najdemo pod Neuvrščeno in Prislov.

4. Ogledamo si besedo *lastnik* in razmislimo o besedni zvezi *lastnik + koga ali česa.* Naštejemo nekaj primerov. Odpremo zavihek Okolica in vnesemo besedo *lastnik* + prvo mesto na desni. Ogledamo si rezultate in spoznamo pojem **kolokator** in različne možnost razvrščanja le-teh. Kolokatorje filtriramo na samostalnike in uredimo glede na pojavitve v okolici.

5. Kdo najde samostalnik, ki se v korpusu najpogosteje pojavlja? Kot zanimivost si ogledamo besedni oblak z najpogostejšimi samostalniki iz korpusa Kres.

6. Kako blizu so bili naši odgovori korpusnim podatkom? Zakaj se je naša **jezikovna intuicija** v nekaterih primerih izkazala za ustrezno, v drugih ne?

Korpus GOS: http://www.korpus-gos.net

1. Odpremo korpus GOS in vtipkamo v iskalno okence besedo *recimo.* Ogledamo si rezultate v konkordančniku in spoznamo razlike korpusa GOS glede na Kres:

- možnost poslušanja posnetka (poslušamo nekaj primerov),
- podatki o besedilih so drugačni (*kaj pravte recimo na to vse vi ste tudi podjetnica* – ogledamo si prekrivanje lastnih imen in oznako za nerazumljivo),
- razlike v filtrih, poskušamo filtrirati, npr. *Maribor, nejavni nezasebni*. Kakšne vrste komunikacije se najdejo v tej vrsti oznake?

2. Poiščemo v korpusu besedo *ful* in si ogledamo filtre. Kaj lahko ugotovimo glede tipa govora, regije snemanja, spol, starost … Lahko glede na podatke zaključimo, da ženske uporabljajo besedo *ful* pogosteje kot moški? Za pomoč pri odločitvi si ogledamo podatke O korpusu.

3. Ponovimo, da je korpus GOS transkribiran na dveh nivojih. Odpremo zavihek Seznam in izberemo Iskanje po standardiziranem zapisu, iščemo besedo *lahko.* Ogledamo si rezultate in vsak poišče obliko in posluša posnetke, ki so najbližje njegovemu narečju ali ki se mu zdijo najbolj zanimivi.

Od konkordance do kolokacije – drugi del

Korpus Janes:

http://nl.ijs.si/noske/janes04.cgi/first_form?corpname=janes.04

1. Odpremo korpus Janes in vnesemo v iskalno okence besedo *valjda.* Ogledamo si rezultate v konkordančniku in spoznamo značilnosti vmesnika:

- konkordančni niz je podoben, na levi imamo nekaj osnovnih **metainformacij** o viru besedila, npr. da gre z tvit in kdo je avtor,
- klik na konkordančno jedro odpre sobesedilo, klik na informacije na levi dodatne metapodatke,
- možnost, da si ogledujemo zadetke v obliki povedi (*Možnosti prikaza > Stavek*),
- možnost, da podatke sortiramo – ogledamo si funkcijo *Premešaj,* ki premeša zadetke, da npr. niso na začetku samo tviti,
- omenimo možnost, da vzorčimo in filtriramo konkordančni niz,
- ogledamo si seznam oblik besede (*Frekvenca > Oblike niza*),
- ogledamo si seznam izvornih dokumentov (*Frekvenca -> Dokumenti*)
- ogledamo si možnosti izdelave seznama kolokatorjev: atribute *word,* v razponu *1* do *1.* Uredimo zadetke po frekvenci in se pogovorimo o rezultatih.

2. Samostojno delo 1: primerjava besed *neumen* in *glup* v korpusu Janes.

- Raziskovanje poteka v parih.
- Eden od dijakov poišče v korpusu Janes besedo *glup,* drugi pa besedo *neumen.*
- Ugotovita, kako pogosto se vsaka od besed pojavlja in postavita hipotezo, zakaj je tako,
- Izdelata seznam kolokatorjev na mestu desno tik ob besedi in si ogledata prvih 100 (dve strani) rezultatov,
- Primerjata oba seznama kolokatorjev in po potrebi posamične konkordance (s klikom na P pred kolokatorjem): kakšno besedišče se pojavlja? Kako pogosti so kolokatorji?

- Katere ugotovitve lahko sklenemo iz podatkov? So podatki potrdili ali ovrgli hipotezo?
- (V sredo bomo spoznali orodje *Primerjalne skice*, ki olajša tovrstne primerjave med besedami.)

3. Samostojno delo 2: različni žanri korpusa Janes.

- Delo poteka v štirih skupinah.
- Vsaka od skupin dobi enega od žanrov: *tvit, blogi, forumi* ali *komentarji* in v konkordančniku za delo izbere ustrezen podkorpus.
- Izberemo besedo ali besedno zvezo, za katero predvidevamo, da se bo pojavljala v vseh žanrih (npr. *itak*).
- Vsaka skupina v svojem podkorpusu naredi naključen vzorec stotih konkordanc.
- Vsaka skupina pregleda svoje konkordance (stavčni pogled) in zabeleži čim več zanimivosti v zvezi z jezikom v njih, npr. ali se pojavljajo posebni znaki in če da, v kakšni funkciji so, kako se uporabljajo ločila, kako se besede zapisujejo, se pojavljajo tujejezične besede ali zveze, sleng, ali katera od besed dobiva nov pomen, kakšen je odnos avtorjev besedila do vsebine, ki jo sporočajo …
- Skupine primerjajo rezultate in ugotovimo, katere značilnosti se pojavljajo v različnih žanrih, katere pa so značilne za posamezen žanr.

Korpus Šolar: www.**korpus-solar**.net/

1. Odpremo korpus Šolar in vnesemo v iskalno okence kot napako besedo *otrok.* Ogledamo si rezultate v konkordančniku in spoznamo razlike vmesnika:

- mogoče je iskati po jezikovnih napakah učencev in popravkih učiteljev,
- napake in popravki so izpisani v konkordančnem nizu in v razširjenem kontekstu,
- možnosti urejanja rezultatov so primerljive, ogledamo si podatke *Oblike niza, Dokumenti.*

2. Razmislimo in naštejemo nekaj napak, ki se nam zdijo tipične za šolska besedila. Ogledamo si možnosti iskanja s pomočjo *Oznake napake*, npr. napake na ravni besedišča, če je čas, še seznam *Oblike niza* in posamezne primer (npr. *in, Hamlet*).

Za obnovitev znanja in dodatne ideje glede uporabe korpusov lahko obiščete spletno stran **Portal jezikovnih virov (viri.trojina.si).** *Janes* sicer še ni predstavljen, so pa na voljo videopredstavitve korpusov *Gigafida* (ki ima enak konkordančnik kot *Kres*), *GOS* in *Šolar*.