GORANKA BLAGUS BARTOLEC i IVANA MATAS IVANKOVIĆ

Kad nam korpus ispunjava želje*

Računalo je danas postalo sredstvo za rad i bez njega više gotovo da i ne možemo zamisliti svakodnevicu. Tako su i jezična istraživanja postala nezamisliva bez pomoći računala. Primarna građa koja nam pritom služi jest računalni korpus – skup strojno čitljivih tekstova nekoga jezika sastavljen po određenome kriteriju. Korpus služi u pretraživanju, prikupljanju i obradi podataka na temelju kojih ćemo donijeti zaključak o pojavi koju istražujemo ili potvrditi hipotezu utemeljenu na drugim izvorima te će nam pomoći u gramatičkim, leksikografskim i drugim jezikoslovnim istraživanjima. Hrvatski jezik pretraživ je na trima računalnim korpusima: na *Hrvatskoj jezičnoj riznici* Instituta za hrvatski jezik i jezikoslovlje (http://riznica.ihjj.hr/), na *Hrvatskome nacionalnom korpusu* (http://filip.ffzg.hr/cgi-bin/run.cgi/first_form) te na *Hrvatskome mrežnom korpusu* hrWaC v2.2 (http://nlp.ffzg.hr/resources/corpora/hrwac/). Ovdje će se ukratko opisati rad na hrWaC-u koji sadržava gotovo 1,4 milijarde pojavnica i najveći je korpus hrvatskoga jezika, a obuhvaća tekstove prikupljene s novinskih portala, foruma i mrežnih stranica službenih organizacija te donosi jezične potvrde koje najvećim dijelom pripadaju publicističkomu, razgovornomu i administrativnomu stilu hrvatskoga standardnog jezika. hrWaC je obrađen i pretraživ u javno dostupnoj programskoj podršci za rad na korpusu NoSketch Engine.

Jednostavna pretraga

Osnovni je i najjednostavniji oblik pretrage korpusa da u okvir *Jednostavna pretraga* upišemo riječ koja nas zanima. Današnji korpusi primjenjuju razvijene pozadinske alate koji omogućuju tagiranje (engl. *tagging*), tj. identifikaciju vrste riječi i morfoloških oblika pojedinih riječi, te parsiranje (engl. *parsing*), tj. rečenično raščlanjivanje pojavnica. Na primjeru imenice *pjesma* prikazat ćemo različite mogućnosti korpusne pretrage. Ako upišemo imensku riječ u nominativu, korpus je prepoznaje kao lemu, dakle tagiranu vrijednost, što znači da će osim kanonskoga oblika pretragom pronaći i sve druge njezine oblike. No, ako u tražilicu upišemo neki padežni oblik (*pjesme*, *pjesmu*, *pjesmi*) ili dvije riječi, npr. *lijepa pjesma*, prikazat će se samo primjeri sa zadanim oblicima. Pritisnemo li u lijevome izborniku opciju KWIC (engl. *key word in context*), dobit ćemo prikaz konkordancija, odnosno popis riječi u okolini, pri čemu je tražena riječ u sredini, a oko nje se pokazuju riječi koje ulaze u zadani opseg broja znakova (npr. 80 znakova s lijeve i s desne strane). Ako odaberemo opciju *Sentence*, dobit ćemo primjere rečenica

* Rad je izrađen u okviru istraživačkoga projekta *Hrvatski mrežni rječnik – Mrežnik* (IP IP-2016-06-214), koji financira Hrvatska zaklada za znanost.

1. slika

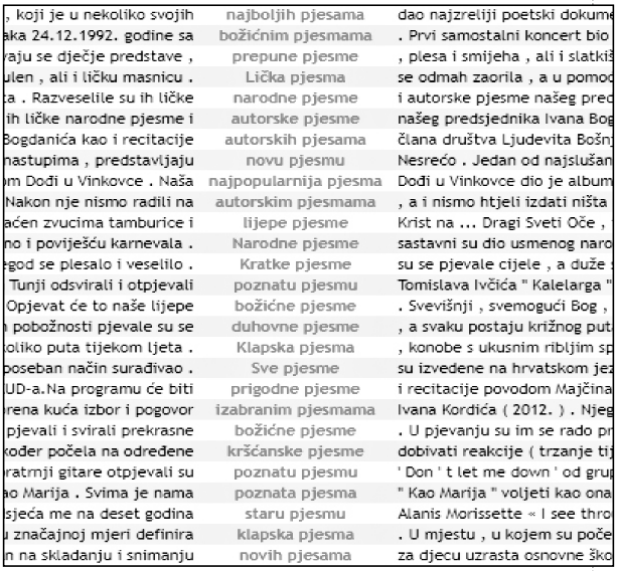
u kojima se nalazi tražena riječ, a zadana riječ neće se prikazivati u središnjemu stupcu.

Na 1. slici prikazani su rezultati pretrage leme *pjesma* s opcijom KWIC.

Dobiveni rezultati izraženi su i brojčano te u lijevome gornjem kutu čitamo podatak da se lema *pjesma* u korpusu pojavljuje u 350,167 primjera (što je prosječno 250,5 primjera na njih milijun) i taj nam podatak omogućuje elementarnu statističku usporedbu. Primjerice, lema *slika* pojavljuje se 460,455 puta (329,4 puta na milijun), pa možemo zaključiti da se pojavljuje češće od leme *pjesma.*

No riječ nam je mnogo zanimljivija kao dio veće cjeline, kao dio sintagmatskoga ili sintaktičkoga ustrojstva, a korpus nam uz malo znanja i malo igre omogućuje da dođemo do različitih podataka o nizovima riječi i jezičnim strukturama.

Pretraga s pomoću regularnih izraza



, koji je u nekoliko svojih	najboljih pjesama	dao najzreliji poetski dokume
aka 24.12.1992. godine sa	božićnim pjesmama	. Prvi samostalni koncert bio
vaju se dječje predstave ,	prepune pjesme	, plesa i smijeha , ali i slatkiš
ulen , ali i ličku masnicu .	Lička pjesma	se odmah zaorila , a u pomoc
a . Razveselile su ih ličke	narodne pjesme	i autorske pjesme našeg prec
ih ličke narodne pjesme i	autorske pjesme	našeg predsjednika Ivana Bog
Bogdanića kao i recitacije	autorskih pjesama	člana društva Ljudevita Bošn
nastupima , predstavljaju	novu pjesmu	Nesrećo . Jedan od najslušan
m Dođi u Vinkovce . Naša	najpopularnija pjesma	Dođi u Vinkovce dio je album
Nakon nje nismo radili na	autorskim pjesmama	, a i nismo htjeli izdati ništa
aćen zvucima tamburice i	lijepe pjesme	Krist na ... Dragi Sveti Oče ,
no i poviješću karnevala .	Narodne pjesme	sastavni su dio usmenog naro
god se plesalo i veselilo .	Kratke pjesme	su se pjevale cijele , a duže
Tunji odsvirali i otpjevali	poznatu pjesmu	Tomislava Ivčića " Kalelarga "
Opjevat će to naše lijepe	božićne pjesme	. Svevišnji , svemogući Bog ,
pobožnosti pjevale su se	duhovne pjesme	, a svaku postaju križnog put
oliko puta tijekom ljeta .	Klapska pjesma	, konobe s ukusnim ribljim sp
poseban način suradivao .	Sve pjesme	su izvedene na hrvatskom jez
UD-a.Na programu će biti	prigodne pjesme	i recitacije povodom Majčina
rena kuća izbor i pogovor	izabranim pjesmama	Ivana Kordića (2012.) . Njeg
pjevali i svirali prekrasne	božićne pjesme	. U pjevanju su im se rado pr
koder počela na određene	kršćanske pjesme	dobivati reakcije (trzanje tij
ratrnji gitare otpjevali su	poznatu pjesmu	' Don ' t let me down ' od grup
ao Marija . Svima je nama	poznata pjesma	" Kao Marija " voljeti kao ona
sjeća me na deset godina	staru pjesmu	Alanis Morissette « I see thro
u značajnoj mjeri definira	klapska pjesma	. U mjestu , u kojem su poče
n na skladanju i snimanju	novih pjesama	za djecu uzrasta osnovne ško

2. slika

Za složenija korpusna pretraživanja upiti se upisuju u okvir *CQL* (*Corpus Query Language*), koji se pojavljuje pritiskom na polje *Query types*. On nam pomaže da tražimo ciljane gramatičke ili leksičke uzorke koje ne možemo postaviti u jednostavnoj pretrazi. Za tu je pretragu potrebno i poznavanje regularnih izraza, kojima definiramo uzorke potrebne za pretraživanje teksta. To su posebni simboli kojima opisujemo traženi skup bez potrebe za nabrajanjem svih elemenata skupa. Točka je, primjerice, simbol kojim se pronalazi bilo koji znak, a zvjezdica označuje pojavljivanje prethodnoga izraza 0, 1 ili bilo koji veći broj puta. Regularni izraz upisuje se u navodnike. Tako, upišemo li u pretragu "na.*", dobit ćemo sve primjere u kojima se nalazi *na*, na početku neke riječi ili kao samostalna riječ.

U pretraživanju niza riječi regularne izraze upisujemo u uglate zagrade (koje se dobiju pritiskom na tipke AltGr i F te AltGr i G). Za svaki element u takvu višerječnome nizu uvjeti mogu biti drukčiji. Primjerice, ako nas zanimaju primjeri s oblikom riječi onako kako smo ga upisali, u regularni izraz upisat ćemo *word*; ako nas zanimaju primjeri s riječju u svim oblicima u kojima se pojavljuje, kao kriterij pretrage poslužit će *lemma*, a ako nam trebaju sve riječi s određenim gramatičkim oznakama, kao kriterij ćemo postaviti *tag*. Dakle, ako želimo saznati kakva pjesma može biti, tj. koji sve pridjevi dolaze ispred riječi *pjesma*, u okvir *CQL* upisat ćemo [tag="A.*"][lemma="pjesma"]. Kao rezultat pretrage dobit ćemo sve primjere s nizom od dviju riječi, pri čemu je prva pridjev, a druga je riječ *pjesma* u bilo kojemu padežnom obliku. (Popis svih kratica upotrijebljenih u obradi korpusa dobit će se pritiskom na *Popis oznaka* ispod polja *CQL*, dok se pregledniji popis tagova može naći i u lijevome izborniku pritiskom na *Corpus Info*.). Rezultati te pretrage prikazani su na 2. slici.

Ako nas zanima što se s pjesmom radi, tj. na koje se sve glagole nadovezuje riječ *pjesma*, postavit ćemo pretragu [tag="V.*"][lemma="pjesma"], a ako želimo izdvojiti samo punoznačne glagole (jer se u velikome broju primjera ispred riječi *pjesma* pojavljuje pomoćni glagol), pretraga će glasiti [tag="Vm.*"][lemma="pjesma"] te ćemo dobiti prikaz kao na 3. slici.

i ZRSovci su 2008. godine	snimili pjesmu	i spot posvećenu Boži Težaku
tične Poljakinje spontane	zapjevale pjesmu	slavljeniku . U organizaciji tu
evnik , spomenar , slikaj ,	piši pjesme	... Počni se baviti vrtlarstvom
OBZORA Ovaj čovjek nije	shvatio pjesmu	. Ne shvaća je ni danas . peta
kako dvogodišnjak pjeva i	svira pjesmu	od Beatlesa Opis videa : Dvog
i ophodarima čim sa šora	odjekne pjesma	Otvarajte kapiju i vrata , evo
n albumu i turneji Johnny	izvodi pjesme	koje su obilježile njegovu kar
nastanak kampova javno	skladao pjesmu	" Konju jedan - mulo jedna ,
do 16:45 sati . " Kome ja	poklonim pjesmu	i spot " 100 godina BOŽE " i m
raljevstvu , a drugi CD će	sadržavati pjesme	snimljene po cijelom svijetu
koji su okupljene za kraj	pozdravili pjesmom	On dolazi . Došao nam je Bož
rija Bitunjac . Priredba je	završila pjesmom	Svi slavimo , a svi prisutni su
3:29:15) dezulovic i lucic	sklepali pjesmu	ivancicu buahahahah TheJa (
13. , 13:20:19) na uvrede	odgovarajte pjesmom	, primijetio je milanovic na p
i će na svom bingo listiću	križati pjesme	koje su čuli , a prvi koji popu
ida o medvjedu " u kojem	pjeva pjesmu	Phil Collinsa , a koju je u orig
ragedija , te za tu priliku	snimaju pjesmu	Čudesan svijet . Suradivala je
vu potrebnu tehnologiju ,	počinje pjesma	, Sinatra daje sve od sebe , p
izvaditi par stotina eura ,	snimiti pjesmu	i misliti kako si zvijezda . Pro
lici uvježbali i na priredbi	otpjevali pjesmu	Vjeruj u ljubav . Naš ovogodi
šlogodišnjih Kvarnerića te	recitirali pjesme	Stojana Tobreca Tote . Da ne
dine ste na CMC festivalu	pjevali pjesmu	Nije muško , a prije godinu d
e ostaju bez nasljedstva ,	zamire pjesma	radosnica i nestaje stvaralašt
ačice krhkih konzonanata	skovati pjesmu	. Mnogi su ga izvršili majstors
biljegom antistaljinizma ,	začinio pjesmom	: Lažu Rusi , jebli svoju majk
ojke su u svom svôm jadu	pjevale pjesmu	koja je kazivala više od na pr
i rat , u kojem se također	pjevala pjesma	čiji je zvon utjeravao strah u
zbu . Trebalo je stvoriti i	njegovati pjesmu	na osnovi hrvatskoga narodno
albansku riječ KANG , što	znači pjesma	, veselje , zabava Narod smat
ati ostale . Glavni pjevač	započinje pjesmu	i predvodi ostale u pjevanju
o pjevači natječu , tko će	započeti pjesmu	. Visina i brzina napjeva zavis

3. slika

pa je tako ova godina	bila iznad očekivanja	, što se tiče dramske s
ota . Ako strelicu miša	dovedete iznad lokota	, pojavit će se natpis "
, brijete se ili bilo što	radite iznad umivaonika	, nema smisla da teče
. 8. Ivanov rov Rov se	nalazio iznad Sokolice	, a staza je vodila blizu
varke i da naši jarboli	strše iznad rijeke	da je naše korijenje to
macija : roll bar mora	prolaziti iznad glave	vozača / suvozača i im
okazalo se da su želje	bile iznad mogućnosti	. Da bi radnici HŽ Infra
nine , čije se ruševine	nalaze iznad sela	Dragović . Značenje im
ko kabelska instalacija	ide iznad kutije	, kutija se može stavit
u obliku u kojem ste ih	naveli iznad napomene	(bez spojnice) nisu u
dnu pletaru , postavio	je iznad roja	i sjetio se tehnike pok
ore u oblacima trese i	propada iznad Mediterana	? Super kaže Vinko , a
voda u hrvatskoj ipak	je iznad kvalitete	vode u razvijenim eur
zajedničko to da žele	ostati iznad politike	. Mehanizmi konstituci
ća koji sa 6 - 7 metara	puca iznad grede	. Za malo je 13. minut
tuka koja daje , uvijek	je iznad ruke	koja prima . " - iz Kale
fiju uključim i ono što	je iznad slapa	, zajedno s nebom ispu
mirena Puležanka koja	stanuje iznad marketa	, svega pedesetak met
eni da naučava i da se	uzdiže iznad muškaraca	. Prvo je bio stvoren A
koja je u tom trenutku	bila iznad glava	mjerila se u milijunima
na leđa kako bi mogla	poletjeti iznad vode	. Voljela je taj osjećaj
uje zakonodavcima da	je iznad ustavâ	, iznad demokracije ,
apon punjenja iskazan	je iznad ikone	koja daje preporuku o
gnite ih u zrak tako da	budu iznad razine	trupa , to će jednostav
e stranice . Unaprijed	zahvaljujem Iznad bazena	pored Hotela Ivan u tu
porast temperatura ne	krije iznad površine	mora , nego neposredn
ovora kao propisa koji	je iznad zakona	određuje superiornost
na vjerskoj dogmatici	postavila iznad zakona	i da se , k tome , još o
vori Gospodin . Visoko	je iznad zemlje	nebo : tako su visoko p

4. slika

Na prvome je mjestu primjer u kojemu se pojavljuje niz *snimili pjesmu*, što nas može potaknuti da se zapitamo koje još imenice mogu doći uz glagol *snimiti*. U korpusu ćemo postaviti pretragu [lemma="snimiti"] [tag="N.*"] te ćemo vidjeti da se može snimiti *probleme*, *pjesmu*, *video*, *EEG*, *zapise*, *podatke*, *predavanje*...

Niz koji se pretražuje može biti i veći. Možda nas zanima koje glagole i imenice povezuje prijedlog *iznad*. Za to nam treba regularni izraz [tag="V.*"] [lemma="iznad"][tag="N.*"] te ćemo dobiti primjere kao na 4. slici.

Već nekoliko svladanih regularnih izraza omogućuje nam da brzo dođemo do velikoga broja primjera za određenu sintaktičku strukturu ili niz riječi. Zahtjevnije pretrage uvjetuju i složenije upite. Uz malu pomoć sa strane (popis regularnih izraza lako se može pronaći na internetu, npr. na adresi https://www.sketchengine.co.uk/documentation/corpus-querying/) mogućnosti pretrage gotovo su neograničene, a korpus time postaje neiscrpno vrelo jezikoslovnih podataka u skladu sa željama i potrebama korisnika.