MAJA MATIJEVIĆ

Riječ, stvar, obitelj... i što još? – Zna li korpus što je *i*-sklonidba?[1]

Hrvatski se jezik učenicima kojima on nije materinski uglavnom predstavlja u kategorijama, pa se o imenicama već u početnim lekcijama govori kao o riječima koje se pojavljuju u trima rodovima – muškome, ženskome i srednjemu. Ti se rodovi najočitije razlikuju prema nominativnome gramatičkom morfu, i to tako da imenice muškoga roda ne završavaju samoglasnikom (osim imenica stranoga podrijetla, tipa *auto*, *avokado* i sl.), imenice ženskoga roda završavaju na *-a*, a imenice srednjega roda na *-o* ili *-e* (ovisno o tome nalazi li se prije gramatičkoga morfa „neki čudan hrvatski glas" poput *č*, *ć*, *lj*, *nj* i sl.).

> Gramatički je morf u hrvatskome završni dio riječi koji služi za tvorbu oblika iste riječi (npr. učiteljic**a**, učiteljic**e**, učiteljic**om**) i iz kojega se može iščitati rod, broj i padež imenica. U hrvatskim se gramatikama gramatički morf obično naziva *nastavak*, ali *nastavak* osim toga može služiti i za tvorbu novih riječi (npr. učitelj**ic**a, učitelj**ev**, učitelj**ski**), pa značenje tih dvaju naziva treba jasno razgraničiti.

Iz te se podjele izdvajaju imenice *i*-sklonidbe, tj. imenice koje ne završavaju na *-a*, a ipak su ženskoga roda. Inojezični ih učenici moraju naučiti napamet, a uvježbavanje se temelji na mnogim primjerima kako bi učenici, osim usvajanja konkretnih riječi, osvijestili i to da pridjev uz imenicu i dalje zadržava uobičajen oblik kao u ostalih imenica ženskoga roda (pa nije **dobar večer*, nego *dobra večer* kao što je i *dobra žena*). Nastavnik hrvatskoga kao drugoga ili stranoga jezika koji sastavlja zadatke pritom se najčešće oslanja na vlastitu intuiciju izvornoga govornika jer iscrpan popis imenica *i*-sklonidbe u hrvatskim gramatikama i priručnicima za hrvatski kao drugi i strani jezik ne postoji. Popis koji se oslanja na rječnike i gramatike objavio je Ivan Marković u časopisu *Lahor* 2007. godine u radu *Do kosti: Imenice hrvatske i-sklonidbe*. Ipak, i bez intuicije i popisa, nastavnik do prigodnih primjera može vrlo lako doći i pretraživanjem korpusa.

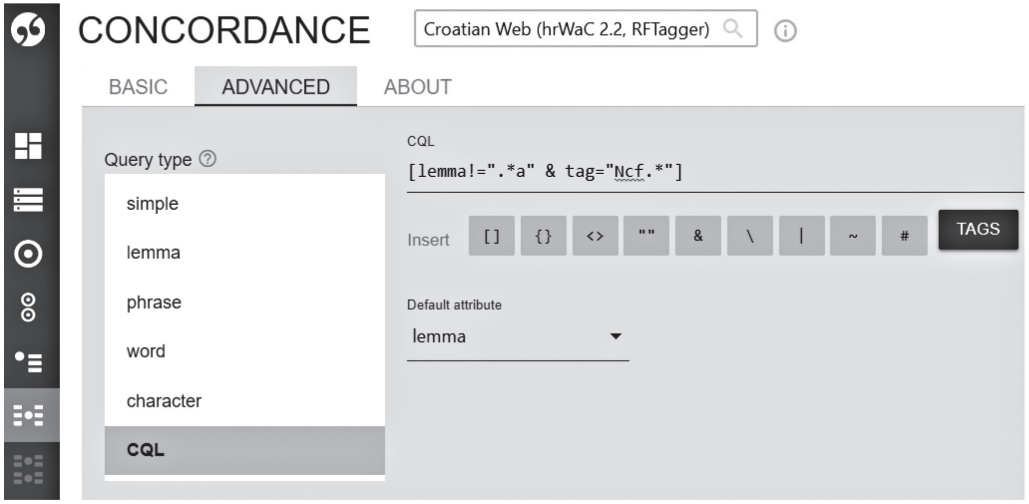
Korpusi su danas vrijedan jezični izvor kojim se koriste gotovo svi koji se bave jezikom: lingvisti, leksikografi, terminolozi, nastavnici… Njihova je najveća prednost u tome što su označeni morfosintaktičkim oznakama kao što su *N* za imenicu, *m* za muški rod, *g* za genitiv, pa upiti osim jednostavnih (npr. pretraživanje rečenica u kojima se pojavljuje

[1] Rad je izrađen u okviru istraživačkoga projekta *Hrvatski mrežni rječnik – Mrežnik* (IP-2016-06-214), koji u cijelosti financira Hrvatska zaklada za znanost i koji se provodi u Institutu za hrvatski jezik i jezikoslovlje.

lak) mogu biti i složeniji (npr. pretraživanje rečenica u kojima se pojavljuje *lak* koji je označen kao pridjev). Hrvatski jezik trenutačno se može pretraživati u trima korpusima: *Hrvatskoj jezičnoj riznici*, *Hrvatskome nacionalnom korpusu* te *Hrvatskome mrežnom korpusu – hrWaC-u*. Potonji je korpus najveći, sastoji se od svih vrsta tekstova koji su bili objavljeni na internetskim stranicama s *.hr* domenom, pretraživ je korpusnim alatom *Sketch Engine* te je javno dostupan akademskoj zajednici. Složeni se upiti formuliraju na *Sketch Engineu* unutar polja *CQL* (*Corpus Query Language*), što zapravo označuje upitni jezik i točno određena pravila na koja korisnik treba prevesti svoje zahtjeve izražene u prirodnome jeziku. Osnovna su pravila ta da se upiti za pojedinu riječ koja se želi dobiti postavljaju unutar uglatih zagrada ([riječ]), a ostali uvjeti unutar zagrada imaju sintaksu, primjerice, *[riječ/lema="lak" & oznaka="imenica"]*. Upiše li se unutar zagrade *word="lak"*, svi će rezultati biti isključivo *lak*. Upiše li se pak *lemma="lak"*, rezultati će biti u svim padežima i, u ovome slučaju, odnosit će se na imenicu *lak* i na pridjev *lak*. Ako se pak želi specificirati rezultate, u upit se kao uvjet može dodati i oznaka (*tag*). Nadopuni li se prethodni upit oznakom za imenicu, glasit će *[lemma="lak" & tag="N.*"]*. U ovome se upitu pojavljuju i dva nova znaka: . (što označava bilo koji znak) i * (što označava ponavljanje prethodnoga znaka bilo koliko puta). Uporaba je tih dvaju znakova na ovome mjestu nužna jer je imenica *lak* u nominativu unutar sustava označena kao *Ncmsn* (***N**oun* za imenicu, ***c**ommon* za opću imenicu, ***m*** za muški rod, ***s*** za jedninu i ***n*** za nominativ). Popis svih oznaka upotrijebljenih za hrvatski jezik nalazi se na adresi https://www.sketchengine.eu/multext-east-croatian-part-of-speech-tagset/. Budući da je u ovome slučaju razlikovno već to da je riječ o imenici, a ne pridjevu (koji bi bio označen s *A* za *adjective*), dovoljno je napisati da je uvjet da je oznaka *N* (a svi ostali znakovi koji se pojavljuju u oznaci nisu važni). Ostali se korisni znakovi mogu ilustrirati primjerima, npr. *[lemma="lak" & tag!= "N.* "]* daje sve oblike riječi *lak* koje nisu (!) označene kao imenica, *[lemma="lak" & tag="N.* "][][lemma="nokat"]* daje sve oblike riječi *lak* koja je označena kao imenica nakon koje slijedi bilo koja riječ (prazne uglate zagrade) te sve oblike riječi *nokat*. Rezultati su ovoga pretraživanja, primjerice, *lak za nokte* i *laka na noktima*.

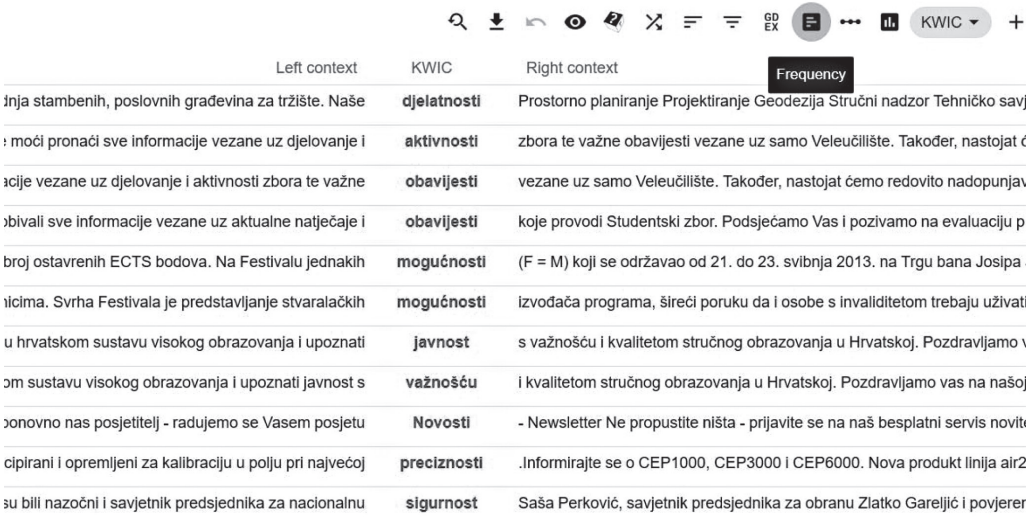
Iako se ovakva sintaksa može činiti složenom, ona se može vrlo brzo usvojiti i korisnik se vrlo brzo može naviknuti na formaliziranje prirodnoga jezika. Poželi li, dakle, nastavnik hrvatskoga jezika koji priprema materijale za nastavni sat o *i*-sklonidbi pronaći više primjera, poželjet će iz korpusa izvući sve imenice koje ne završavaju na *-a*, a pripadaju ženskome rodu, odnosno:

Upitni jezik: **[lemma!=".*a" & tag="Ncf.* "]**
Prirodni jezik: izbaci mi sve oblike riječi (lemma) koje ne završavaju (!=) na -a (".*a") i (&) označene su (tag=) kao opće imenice ženskoga roda ("Ncf.*")
U obama slučajevima * znači pojavljivanje bilo kojega znaka bilo koliko puta.



1. slika: Složeni upit za imenice *i*-sklonidbe

Korpus izbacuje nizove rečenica u kojima se pojavljuju riječi koje zadovoljavaju postavljeni uvjet. Ako korisnik potom želi prikaz bez ponavljanja istih primjera, može mišem odabrati frekvenciju u gornjemu desnom kutu te potom frekvencijski popis lema unutar ključne riječi u kontekstu (*KWIC*, *key word in context*).



2. slika: Dobiveni rezultati

Sketch Engine potom razvrstava sve dobivene rezultate prema kanonskome obliku (lemi) i slaže ih prema čestotnosti.

	Lemma	↓ Frequency	Frequency per million	
1	riječ	452,566	321.93	•••
2	stvar	387,350	275.54	•••
3	obitelj	279,576	198.87	•••
4	mogućnost	268,192	190.78	•••
5	ljubav	257,117	182.90	•••
6	pomoć	229,816	163.48	•••
7	vlast	199,477	141.90	•••
8	vrijednost	197,614	140.57	•••
9	aktivnost	180,439	128.35	•••
10	bolest	175,817	125.07	•••
11	povijest	169,899	120.86	•••
12	javnost	152,576	108.53	•••
13	momčad	144,392	102.71	•••
14	smrt	127,919	90.99	•••

3. slika: Dobiveni rezultati bez ponavljanja, složeni prema čestotnosti

S vrlo malo izraza korpusi se mogu pretraživati s obzirom na različite kriterije, a korisnik lako može prilagoditi pretraživanje svojim potrebama. Ako mu se učini da među rezultatima ima previše riječi *i*-sklonidbe koje završavaju na *-st* ili *-ad*, može pretraživati tako da mu korpus takve riječi ne izbacuje uvrstivši u oble zagrade sve nastavke koje ne želi i razdvojivši ih ravnom crtom (*[lemma!=".*(a|st|ad)" & tag="Ncf.*"]*). Ovaj upit dat će imenice *i*-sklonidbe s raznovrsnijim tvorbenim uzorkom.

Na dobivene se rezultate izvorni govornik može osloniti, ali pri ovakvome pretraživanju svakako treba uzeti u obzir i to da je *hrWaC* skup tekstova koji ni na koji način nisu probrani i u kojemu su, osim toga, morfosintaktičke oznake dodane automatski. Zbog toga se među rezultatima može pronaći *čast* i *ćast* i *cast*; *povijest* i *povjest*; *raj*, *lijep*, *prepun* i dr. U radu s korpusima, dakle, ipak treba biti svjestan toga da *hrWaC* nije korpus hrvatskoga standardnog jezika te da nisu sve unesene oznake točne, zbog čega se mogu pojaviti žargonizmi ili dijalektalizmi koji pritom nisu dobro označeni. No, brojni željeni rezultati koji su dobiveni pokazuju kako svojevrsno prevođenje vlastitih misli na jezik razumljiv računalu može skratiti posao i inspirirati nastavnika, i sve to – u dva klika mišem.