MARIO MARKOVIĆ*,
JOSIP MIHALJEVIĆ,
MILICA MIHALJEVIĆ

Kako pronaći jezikoslovni naziv[1]

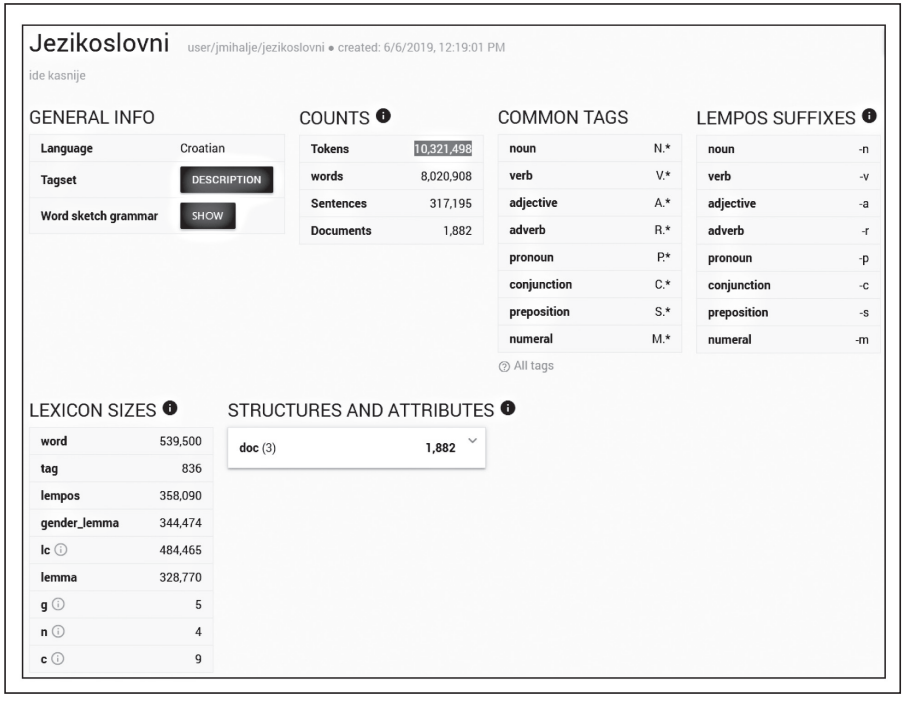
Projekt *Hrvatsko jezikoslovno nazivlje – Jena* započeo je 24. svibnja 2019. i traje do 23. studenoga 2020. Projekt je pokrenut zbog potrebe ususta vljivanja hrvatskoga jezikoslovnog nazivlja, a provodi se u sklopu programa *Struna*. Temeljni je cilj projekta tijekom projektnoga razdoblja unijeti u bazu *Strune* 1500 naziva sa sinonimima, antonimima, podređenim nazivima, definicijama i istovrijednicama na pet svjetskih jezika: engleskome, njemačkome, francuskome, ruskome i švedskome. Takva baza bit će temelj za daljnje ususta vljivanje jezikoslovnoga nazivlja jer će se i nakon završetka predviđenoga trajanja projekta upotpunjavati dodavanjem novih naziva. Za hrvatski jezik ne postoji specijalizirani korpus jezikoslovnoga nazivlja. Većina se korpusnih istraživanja provodi na korpusima *Hrvatska jezična riznica* (http://riznica.ihjj.hr/index.hr.html) i *Hrvatski jezični korpus* hrWaC (http://nlp.ffzg.hr/resources/corpora/hrwac/). Ti korpusi, međutim, obuhvaćaju jako malo tekstova koji pripadaju znanstvenomu stilu, pa su stoga neprikladni za terminološka istraživanja. Stoga će se, da bi se postigao definirani cilj, izraditi priručni jezikoslovni korpus koji se sastoji od računalno dostupnih i pretraživih izvora. Taj će korpus biti temelj za daljnja proučavanja jezikoslovnoga nazivlja i on će se dopunjavati novim izvorima i nakon završetka trajanja projekta. Cilj je ovoga rada pokazati na koji je način pristupljeno izradi jezikoslovnoga korpusa kako bi se o postojanju i značajkama korpusa informiralo sve one koje zanima jezikoslovno nazivlje, ali i kako bi se pružio model oblikovanja specijaliziranih korpusa, koji je primjenjiv i na druge struke. Dosad je izrađena radna inačica korpusa, koja je dostupna svim članovima projekta *Jena*, a na zahtjev i svim ostalim članovima akademske zajednice koji imaju pristup programu *Sketch Engine* te AAI@EduHr korisnički račun.

Kako bi se izradio jezikoslovni korpus, u program *Sketch Engine* učitane su jezikoslovne knjige i časopisi u formatu .txt. U korpusu se zasad nalaze odabrani časopisi koji se mogu besplatno preuzeti s portala znanstvenih časopisa *Hrčak* te odabrane knjige kojima je nositelj autorskih prava Institut za hrvatski jezik i jezikoslovlje. U ovoj fazi nisu uključene knjige iz povijesti jezika i dijalektologije koje sadržavaju specijalne znakove i

* Mario Marković nastavnik je informatike u Školi za primalje i u Upravnoj školi u Zagrebu, a bio je honorarni suradnik projekta *Jena* koji je radio na izradi jezikoslovnoga korpusa.

[1] Rad je izrađen u okviru istraživačkoga projekta *Hrvatsko jezikoslovno nazivlje – Jena* (Struna-2017-09-05), koji u cijelosti financira Hrvatska zaklada za znanost i koji se provodi u Institutu za hrvatski jezik i jezikoslovlje.

grafiju. Kako bi ovaj korpus postao reprezentativan, trebalo bi uključiti više knjiga drugih izdavača. Korpus sadržava ove časopise: *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, *Hrvatski jezik: znanstveno-popularni časopis za kulturu hrvatskoga jezika*, *FLUMINENSIA: časopis za filološka istraživanja*, *Suvremena lingvistika*, *Folia onomastica Croatica*, *Filologija*, *Jezikoslovlje* i ove knjige: *O umu stručnjaka* (Nahod 2016.), *Hrvatski terminološki priručnik* (Hudeček i Mihaljević 2012.), *Izražavanje prostora i vremena prijedlozima s genitivom u hrvatskom i ruskom jeziku* (Matas Ivanković 2014.), *Ja, Krsto Lučin Dubrovčanin, činim ovi testamenat…* (Lovrić Jović 2015.), *Hrvatski na maturi sa zadatcima za vježbu* (Hudeček i Mihaljević 2016.), *Glagolski vid u hrvatskim gramatikama do 20. stoljeća* (Brlobaš 2007.), *Hrvatska školska gramatika* (Hudeček i Mihaljević 2017.), *Instrumental u hrvatskom jeziku* (Brač 2018.), *Praktični vodič kroz mišljenje i značenje* (Jackendoff 2012.), *Struktura povratnih glagola i konstrukcije sa se u hrvatskome jeziku* (Oraić Rabušić 2018.), *Oblici nebrojivosti u hrvatskom jeziku* (Peti 2004.), *O rodu jezikom i pokoja fraška* (Vidović 2019.), *Izražavanje posljedičnih odnosa u hrvatskome standardnom jeziku* (Vukojević 2008.), *Valencijski rječnik psiholoških glagola u hrvatskome jeziku* (Birtić i sur. 2018.), *Unutarnja struktura odglagolskih imenica u hrvatskome jeziku* (Birtić 2008.). Korpus trenutačno ima 1882 različitih izvora (knjiga i članaka), 1 032 498 pojavnica i 8 020 908 riječi i stalno se dopunjuje.

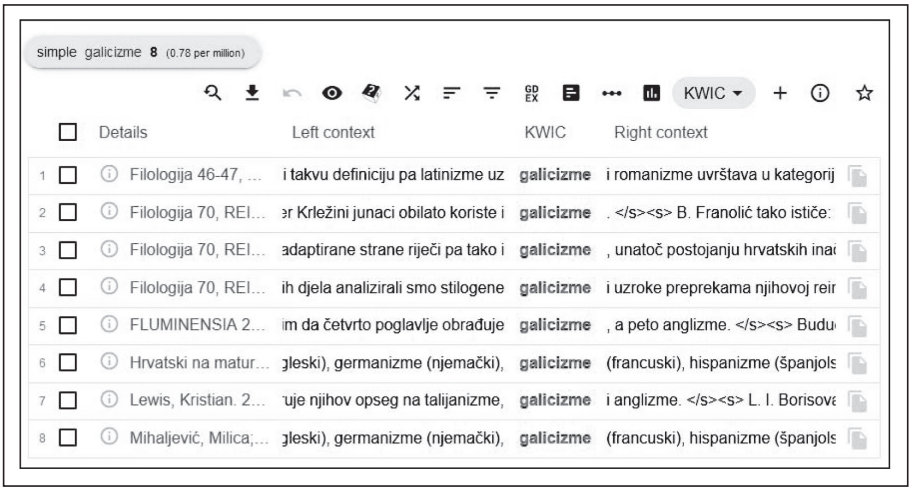


1. slika: Prikaz podataka o *Jezikoslovnome korpusu*

Članci su preuzeti jedan po jedan, pri čemu je svaki članak preimenovan tako da je ime datoteke istovjetno naslovu članka. Datoteke u svojemu imenu osim naslova članka sadržavaju i ime časopisa te godište i broj časopisa. To je napravljeno kako bi korisnik mogao vidjeti izvor potvrde koju pronađe u korpusu. Da bi se više članaka moglo usporedno preimenovati i tako ubrzati posao, upotrijebljen je besplatni alat *Advanced Renamer* (https://www.advancedrenamer.com/, pristupljeno 2. siječnja 2020.). Za knjige su uz naslov dodana i imena autora te godina kad je knjiga izdana. Nažalost, neki stariji brojevi časopisa sadržavaju članke koji su objavljeni kao skenirane slike bez prepoznavanja teksta, pa je bilo potrebno izvršiti optičko prepoznavanje znakova (OCR) s pomoću programa ABBYY FineReader, koji može precizno prepoznati utipkane znakove za hrvatski jezik ako se tekst na slici jasno vidi te je riječ o prepoznatljivome fontu (npr. Times New Roman, Arial, Calibri itd.). Budući da *Sketch Engine* ne može pravilno prikazati slike, tablice, bilješke, brojeve stranica, tekstove zaglavlja itd., bilo je potrebno maknuti ih iz teksta kako bi korpus imao manje pogrešaka u prikazu glavnoga teksta (npr. kod rečenice koja se nastavlja na idućoj stranici broj stranice umetne se usred rečenice jer *Sketch Engine* čita tekst kao jedan odlomak). Problem sa zaglavljem i brojem stranice riješen je tako da su se u dokumentu u PDF-u odredile margine za rezanje zaglavlja i podnožja u programu *Sejda* (https://www.sejda.com/ (pristupljeno 2. siječnja 2020.). Program je omogućio da se usporedno učita više članaka iz određenoga časopisa, od kojih svaki časopis ima određeno područje za zaglavlje i brojeve stranice, te se u grafičkome sučelju moglo precizno odrediti kako će se izrezati članci. Bilješke se nisu rezale jer su drukčije ovisno o količini teksta. Nakon toga se svaki dokument iz PDF-a prebacio u Word (.docx) s pomoću ABBYY FineReadera, koji u procesu prebacivanja ima mogućnost prepoznavanja bilježaka, tablica i slika unutar tekstne strukture. Kod dokumenata u Wordu slike, tablice i bilješke očišćene su s pomoću makronaredbe (pregled makrokoda za pročišćavanje tekstova: https://bit.ly/39uqp1U, pristupljeno 2. siječnja 2020.), koja se jednim pokretanjem izvršava na više dokumenata. Dokumenti u Wordu dodatno su očišćeni te prebačeni u .txt dokumente s pomoću skripte za *Python* (Python skripte za pročišćavanje tekst i prebacivanje u format .txt: https://borna12.gitlab.io/igre-mreznik/kod%20za%20jenu/pretvarac/word%20to%20txt.py, pristupljeno 2. siječnja 2020.), koja je u dokumentima nad kojim je izvršeno optičko prepoznavanje znakova ispravila nekoliko nepoznatih znakova u tekstu (npr. ▯ u *ù*), maknula nepotrebne razmake, nepotrebne spojnice na kraju retka te obrisala popis literature na kraju teksta (početak popisa literature prepoznaje se na temelju ključnih riječi na početku odlomka: *Literatura:*, *Izvori:*, *Vrela:*, *Bibliografija:* itd.). Dodatno je u programu *Notepad++* pregledan svaki .txt dokument kako se ne bi u korpus unijeli tekstovi koji nisu na hrvatskome, te su izbrisani sažetci na stranim jezicima koji su se često nalazili na početku ili kraju teksta. Korpus je još uvijek u demoinačici te još ima nekih uočljivih pogrešaka u tekstu, ali je i

Jezikoslovni korpus jednojezični je (hrvatski) specijalizirani (terminološki) korpus.

u ovome obliku veoma korisno pomagalo za sve koji se bave istraživanjem hrvatskoga jezikoslovnog nazivlja, znanstvenoga stila ili samo žele provjeriti tko je u svojemu radu spomenuo neki naziv. Na 2. slici nalazi se prikaz konkordancije leme *galicizam*.

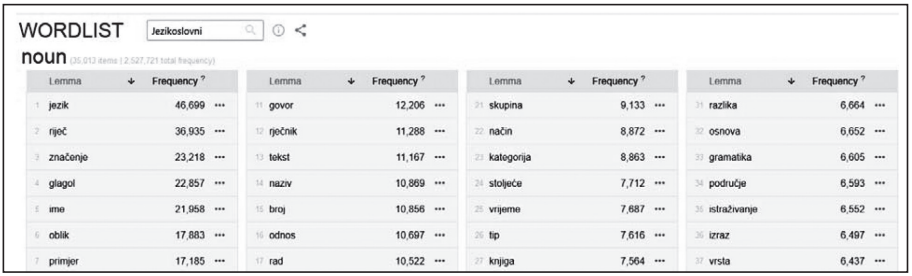


2. slika: Konkordancija leme *galicizam*

Na 3. slici nalazi se prikaz dijela skica riječi za lemu *imenica*.

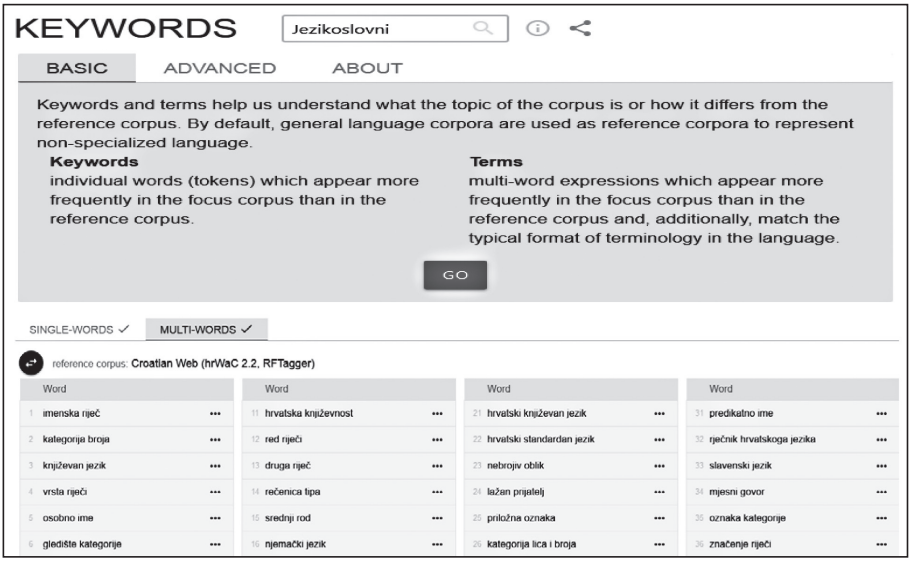
3. slika: Skice riječi za lemu *imenica*

Na 4. slici vidi se popis imenica iz *Jezikoslovnoga korpusa*



4. slika: Popis imenica iz *Jezikoslovnoga korpusa*

Na 5. slici nalazi se popis potencijalnih naziva (ključnih riječi) automatski izlučenih iz *Jezikoslovnoga korpusa*. Nazivi se dobivaju usporedbom frekvencija pojave riječi i višerječnih skupina u jezikoslovnome i u općemu korpusu. Usto se vodi računa i o pretpostavljenoj strukturi naziva.



5. slika: Popis potencijalnih višerječnih naziva (ključnih riječi) iz *Jezikoslovnoga korpusa*