



MARIO MARKOVIĆ*,
JOSIP MIHALJEVIĆ,
MILICA MIHALJEVIĆ

Kako pronaći jezikoslovni naziv¹

Projekt *Hrvatsko jezikoslovno nazivlje – Jena* započeo je 24. svibnja 2019. i traje do 23. studenoga 2020. Projekt je pokrenut zbog potrebe usustavljanja 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 usustavljanje 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*, *FLUMINENSLA: 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.

Jezikoslovnvi

user/jmihalje/jezikoslovnvi • created: 6/6/2019, 12:19:01 PM

ide kasnije

<h4 style="margin: 0;">GENERAL INFO</h4> <p>Language: Croatian</p> <p>Tagset: DESCRIPTION</p> <p>Word sketch grammar: SHOW</p>	<h4 style="margin: 0;">COUNTS</h4> <table style="width: 100%; border-collapse: collapse;"> <tr><td>Tokens</td><td style="text-align: right;">10,321,498</td></tr> <tr><td>words</td><td style="text-align: right;">8,020,908</td></tr> <tr><td>Sentences</td><td style="text-align: right;">317,195</td></tr> <tr><td>Documents</td><td style="text-align: right;">1,882</td></tr> </table>	Tokens	10,321,498	words	8,020,908	Sentences	317,195	Documents	1,882	<h4 style="margin: 0;">COMMON TAGS</h4> <table style="width: 100%; border-collapse: collapse;"> <tr><td>noun</td><td>N.*</td><td>noun</td><td>-n</td></tr> <tr><td>verb</td><td>V.*</td><td>verb</td><td>-v</td></tr> <tr><td>adjective</td><td>A.*</td><td>adjective</td><td>-a</td></tr> <tr><td>adverb</td><td>R.*</td><td>adverb</td><td>-r</td></tr> <tr><td>pronoun</td><td>P.*</td><td>pronoun</td><td>-p</td></tr> <tr><td>conjunction</td><td>C.*</td><td>conjunction</td><td>-c</td></tr> <tr><td>preposition</td><td>S.*</td><td>preposition</td><td>-s</td></tr> <tr><td>numeral</td><td>M.*</td><td>numeral</td><td>-m</td></tr> </table> <p style="font-size: x-small; margin: 0;">All tags</p>	noun	N.*	noun	-n	verb	V.*	verb	-v	adjective	A.*	adjective	-a	adverb	R.*	adverb	-r	pronoun	P.*	pronoun	-p	conjunction	C.*	conjunction	-c	preposition	S.*	preposition	-s	numeral	M.*	numeral	-m	<h4 style="margin: 0;">LEMPOS SUFFIXES</h4> <table style="width: 100%; border-collapse: collapse;"> <tr><td>noun</td><td>-n</td></tr> <tr><td>verb</td><td>-v</td></tr> <tr><td>adjective</td><td>-a</td></tr> <tr><td>adverb</td><td>-r</td></tr> <tr><td>pronoun</td><td>-p</td></tr> <tr><td>conjunction</td><td>-c</td></tr> <tr><td>preposition</td><td>-s</td></tr> <tr><td>numeral</td><td>-m</td></tr> </table>	noun	-n	verb	-v	adjective	-a	adverb	-r	pronoun	-p	conjunction	-c	preposition	-s	numeral	-m
Tokens	10,321,498																																																										
words	8,020,908																																																										
Sentences	317,195																																																										
Documents	1,882																																																										
noun	N.*	noun	-n																																																								
verb	V.*	verb	-v																																																								
adjective	A.*	adjective	-a																																																								
adverb	R.*	adverb	-r																																																								
pronoun	P.*	pronoun	-p																																																								
conjunction	C.*	conjunction	-c																																																								
preposition	S.*	preposition	-s																																																								
numeral	M.*	numeral	-m																																																								
noun	-n																																																										
verb	-v																																																										
adjective	-a																																																										
adverb	-r																																																										
pronoun	-p																																																										
conjunction	-c																																																										
preposition	-s																																																										
numeral	-m																																																										

<h4 style="margin: 0;">LEXICON SIZES</h4> <table style="width: 100%; border-collapse: collapse;"> <tr><td>word</td><td style="text-align: right;">539,500</td></tr> <tr><td>tag</td><td style="text-align: right;">836</td></tr> <tr><td>lempos</td><td style="text-align: right;">358,090</td></tr> <tr><td>gender_lemma</td><td style="text-align: right;">344,474</td></tr> <tr><td>lc</td><td style="text-align: right;">484,465</td></tr> <tr><td>lemma</td><td style="text-align: right;">328,770</td></tr> <tr><td>g</td><td style="text-align: right;">5</td></tr> <tr><td>n</td><td style="text-align: right;">4</td></tr> <tr><td>c</td><td style="text-align: right;">9</td></tr> </table>	word	539,500	tag	836	lempos	358,090	gender_lemma	344,474	lc	484,465	lemma	328,770	g	5	n	4	c	9	<h4 style="margin: 0;">STRUCTURES AND ATTRIBUTES</h4> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="border: 1px solid #ccc; padding: 2px;">doc (3)</td> <td style="text-align: right; padding: 2px;">1,882</td> </tr> </table>	doc (3)	1,882
word	539,500																				
tag	836																				
lempos	358,090																				
gender_lemma	344,474																				
lc	484,465																				
lemma	328,770																				
g	5																				
n	4																				
c	9																				
doc (3)	1,882																				

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 svojem 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,

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

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://bornal2.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. Š u Ÿ), 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

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

simple galicizme 8 (0.78 per million)			
<input type="checkbox"/> Details	Left context	KWIC	Right context
<input type="checkbox"/> 1	Filologija 46-47, ...	i takvu definiciju pa latinizme uz galicizme	i romanizme uvrštava u kategorij
<input type="checkbox"/> 2	Filologija 70, REI...	er Krelžini junaci obilato koriste i galicizme	. </s><s> B. Franolić tako ističe:
<input type="checkbox"/> 3	Filologija 70, REI...	adaptirane strane riječi pa tako i galicizme	, unatoč postojanju hrvatskih inai
<input type="checkbox"/> 4	Filologija 70, REI...	ih djela analizirali smo stilogene galicizme	i uzroke preprekama njihovoj reir
<input type="checkbox"/> 5	FLUMINENSIA 2...	im da četvrto poglavlje obrađuje galicizme	, a peto anglizme. </s><s> Budu
<input type="checkbox"/> 6	Hrvatski na matur...	gleski), germanizme (njemački), galicizme	(francuski), hispanizme (španjols
<input type="checkbox"/> 7	Lewis, Kristian. 2...	tuje njihov opseg na talijanizme, galicizme	i anglizme. </s><s> L. I. Borisov
<input type="checkbox"/> 8	Mihaljević, Milica;...	gleski), germanizme (njemački), galicizme	(francuski), hispanizme (španjols

2. slika: Konkordancija leme *galicizam*

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

kakov?	subjekt_od	u_genitivu-n	n-koga-čega
+ pridjev + imenica	označivati imenice koje označuju	plurali imenice pluralia tantum	sklonidba sklonidbi imenica
opći općih imenica	značiti imenice koje znače	e-vrsta imenica e-vrste	množina množine imenica
zbirni zbirne imenice	završavati imenice koje završavaju na	mreža imenice mreža	tvorba za tvorbu imenica
dogadajan dogadajnih imenica	imati imenice imaju	tip imenice tipa	skupina skupine imenica
apstraktan apstraktne imenice	tvoriti imenice tvore	singulari imenice singularia tantum	deklinacija deklinaciji imenica

3. slika: Skice riječi za lemu *imenica*

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

WORDLIST Jezikoslovni

36,013 items (2,527,721 total frequency)

noun

Lemma	Frequency ?	Lemma	Frequency ?	Lemma	Frequency ?	Lemma	Frequency ?
1 jezik	46,699 ...	11 govor	12,206 ...	21 skupina	9,133 ...	31 razika	6,664 ...
2 riječ	36,935 ...	12 rječnik	11,288 ...	22 način	8,872 ...	32 osnova	6,652 ...
3 značenje	23,218 ...	13 tekst	11,167 ...	23 kategorija	8,863 ...	33 gramatika	6,605 ...
4 glagol	22,857 ...	14 naziv	10,869 ...	24 stoljeće	7,712 ...	34 područje	6,593 ...
5 ime	21,958 ...	15 broj	10,856 ...	25 vrijeme	7,687 ...	35 istraživanje	6,552 ...
6 oblik	17,883 ...	16 odnos	10,697 ...	26 lip	7,616 ...	36 izraz	6,497 ...
7 primjer	17,185 ...	17 rad	10,522 ...	27 knjiga	7,504 ...	37 vrsta	6,437 ...

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.

KEYWORDS Jezikoslovni

BASIC ADVANCED ABOUT

Keywords and terms help us understand what the topic of the corpus is or how it differs from the reference corpus. By default, general language corpora are used as reference corpora to represent non-specialized language.

Keywords
individual words (tokens) which appear more frequently in the focus corpus than in the reference corpus.

Terms
multi-word expressions which appear more frequently in the focus corpus than in the reference corpus and, additionally, match the typical format of terminology in the language.

GO

SINGLE-WORDS ✓ MULTI-WORDS ✓

reference corpus: Croatian Web (hrWac 2.2, RFTagger)

Word	Word	Word	Word
1 imenska riječ	11 hrvatska književnost	21 hrvatsko književan jezik	31 predikatno ime
2 kategorija broja	12 red riječi	22 hrvatski standardan jezik	32 rječnik hrvatskoga jezika
3 književan jezik	13 druga riječ	23 netrojni oblik	33 slavenski jezik
4 vrsta riječi	14 rečenica tipa	24 ližan prijatelj	34 mjesni govor
5 osobno ime	15 srednji rod	25 priložna oznaka	35 oznaka kategorije
6 gladište kategorije	16 njemački jezik	26 kategorija lica i broja	36 značenje riječi

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