

Mario Marković⁶³

Upravna škola Zagreb

Prilaz baruna Filipovića 30, HR-10000 Zagreb

mmarkovi3g@gmail.com

Josip Mihaljević

Institut za hrvatski jezik i jezikoslovlje

Ulica Republike Austrije 16, HR-10000 Zagreb

jmihalj@ihjj.hr

KORPUS *JENE* – JEDNOJEZIČNI TERMINOLOŠKI KORPUS

1. Uvod

Natuknica *korpus* u bazi *Jene* obrađena je ovako (vidi 1. sliku):

korpus

status naziva: naziv

definicija: zbirka tekstova koji su odabrani i skupljeni prema jasnim jezikoslovnim kriterijima radi dobivanja određenoga jezičnog uzorka

vrelo definicije: Pojmovnik, Mrežnik. (mrežno)

projekt: jezikoslovlje

potpodručje: e-leksikografija i korpusno jezikoslovlje

podređeni pojam: govorni korpus; nacionalni korpus; opći korpus; reprezentativni korpus; specijalni korpus; stručni korpus; usporedni korpus

istovrijednica - engleski: corpus

istovrijednica - njemački: Korpus

istovrijednica - francuski: corpus

istovrijednica - ruski: корпус

istovrijednica - švedski: korpus

jezična odrednica: imenica

broj: jednina

rod: muški

napomena: Korpusi se dijele s obzirom na različite kriterije. S obzirom na broj jezika razlikuju se jednojezični, dvojezični i višejezični korpusi. S obzirom na odnos tekstova na različitim jezicima razlikuju se usporedni i usporedivi korpusi, s obzirom na obuhvaćeno razdoblje razlikuju se dijakronijski i sinkronijski korpusi, s obzirom na obuhvaćene tekstove razlikuju se općejezični i specijalizirani korpusi, s obzirom na reprezentativnost korpusa za neki jezik razlikuju se reprezentativni i nerepresentativni korpusi. Korpusi koji uključuju multimedijske elemente nazivaju se multimedijski korpusi. S obzirom na autore tekstova obuhvaćenih korpusom izdvaja se dječji korpus, korpus neizvornih govornika, učenički korpus itd.

1. slika: Natuknica *korpus* u *Jeni*

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

⁶³ Mario Marković nastavnik je informatike u Upravnoj školi u Zagrebu, a bio je honorarni suradnik projekta *Jena* i zajedno s Josipom Mihaljevićem radio je na izradi jezikoslovnoga korpusa.

neprikladni za terminološka istraživanja. Stoga je, da bi se moglo proučavati jezikoslovno nazivlje u okviru projekta *Hrvatsko jezikoslovno nazivlje – Jena*, izrađen priručni kontrolni korpus koji se sastoji od računalno dostupnih i pretraživih izvora.⁶⁴ Taj će korpus biti temelj za daljnja proučavanja jezikoslovnoga nazivlja i on će se dopunjavati novim izvorima u okviru internoga projekta Instituta za hrvatski jezik i jezikoslovlje.

U ovome će se poglavlju pokazati na koji je način pristupljeno izradi korpusa jezikoslovnoga nazivlja 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 korisnički račun AAI@EduHr.

2. Postupak izrade korpusa

Korpus *Jene* specijalizirani je stručni korpus. Na 2. slici nalazi se obrada naziva *stručni korpus* u *Jeni*.

stručni korpus	
status naziva: naziv	
definicija: korpus tekstova određene struke	
vrela definicije: Pojmovnik, Mrežnik (mrežno)	
projekt: jezikoslovlje	
potpodručje: e-leksikografija i korpusno jezikoslovlje	
dopušteni naziv: terminološki korpus	istovrijednica - engleski: domain specific corpus
	istovrijednica - njemački: Fachsprachenkorpus
	istovrijednica - francuski: corpus de langage technique
	istovrijednica - ruski: корпус специальных текстов
	istovrijednica - švedski: fackspråkskorpus
	jezična odrednica: višerječni naziv

2. slika: Obrada naziva *stručni korpus* u *Jeni*

Kako bi se izradio korpus jezikoslovnoga nazivlja, u program *Sketch Engine* (Kilgarriff, Rychlý, Smrz, Tugwell 2004.; *Sketch Engine Guide* 2019.) 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

⁶⁴ O tome je korpusu već pisano u radu Marković, Mihaljević i Mihaljević 2020: 18–22.

jezika i dijalektologije koje sadržavaju specijalne znakove i grafiju. Kako bi *Jezikoslovn* *korpus* postao reprezentativan, trebalo bi uključiti više knjiga drugih izdavača. Časopisi su koji čine korpus *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*, a knjige su koje čine korpus *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 zadacima 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 dr. 2018.), *Unutarnja struktura odglagolskih imenica u hrvatskome jeziku* (Birtić 2008.).

Korpus trenutačno ima 1 882 različitih izvora (knjiga i članaka), 10 32 498 pojavnica i 8 020 908 riječi. Podatci o *Jezikoslovn* *korpusu* prikazani su na 3. slici.

Jezikoslovnji user/jmihalje/jezikoslovnji • created: 6/6/2019, 12:19:01 PM
 ide kasnije

GENERAL INFO		COUNTS ⁱ		COMMON TAGS		LEMPOS SUFFIXES ⁱ	
Language	Croatian	Tokens	10,321,498	noun	N.*	noun	-n
Tagset	DESCRIPTION	words	8,020,908	verb	V.*	verb	-v
Word sketch grammar	SHOW	Sentences	317,195	adjective	A.*	adjective	-a
		Documents	1,882	adverb	R.*	adverb	-r
				pronoun	P.*	pronoun	-p
				conjunction	C.*	conjunction	-c
				preposition	S.*	preposition	-s
				numeral	M.*	numeral	-m
				All tags			

LEXICON SIZES ⁱ		STRUCTURES AND ATTRIBUTES ⁱ	
word	539,500	doc (3)	1,882
tag	836		
lempos	358,090		
gender_lemma	344,474		
lc ⁱ	484,465		
lemma	328,770		
g ⁱ	5		
n ⁱ	4		
c ⁱ	9		

3. slika: Prikaz podataka o *Jezikoslovnome korpusu*

Članci su preuzeti jedan po jedan, pri čemu je svaki članak preimenovan tako da je naziv datoteke odgovara naslovi članka. Datoteke u svojem nazivu 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. Naslovi nekih članaka morali su biti skraćeni zbog ograničenja broja znakova koje može sadržavati naziv datoteke, ali se i dalje može prepoznati o kojemu je članku riječ jer se vidi u kojemu se godištu i broju časopisa nalazi. Također se u nazivima datoteka članaka moralo zamijeniti određene znakove koje sustav ne dopušta (npr. ", ?, :). Podatak o korpusnoj potvrdi iz članka prikazan je na 4. slici.

Display and count metadata

Select the metadata to be displayed in the concordance. Click  to calculate statistics.

Display above lines ? Shorten to 15 characters

I adpozicija kojim se operira u knjizi obuhvaćaju se prijedlozi i **poslijelozi** zajedno (premda izgleda da su određeni samo položajem, ipz

File name   (1)

<input type="checkbox"/> Token number	2965619	
<input type="checkbox"/> Document number	560	
<input type="checkbox"/> doc.wordcount	1062	
<input type="checkbox"/> File ID	file13578033	
<input checked="" type="checkbox"/> File name	FLUMINENSIA 29-1, NOVI PRISTUP HRVATSKIM PRIJEDLOZIMA.txt	

CLOSE SAVE

4. slika: Podatak o korpusnoj potvrdi (članak)

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. Podatak o korpusnoj potvrdi iz knjige prikazan je na 5. slici.

Display and count metadata

Select the metadata to be displayed in the concordance. Click  to calculate statistics.

Display above lines ? Shorten to 15 characters

<s><doc> Osnovni je predmet ove knjige tvorba odglagolskih **imenica** u hrvatskome jeziku. </s></s> Istraživanje se imenica temelji n:

File name   (1)

<input type="checkbox"/> Token number	7	
<input type="checkbox"/> Document number	0	
<input type="checkbox"/> doc.wordcount	59187	
<input type="checkbox"/> File ID	file11527792	
<input checked="" type="checkbox"/> File name	Birtić, Matea. 2008. Unutarnja struktura odglagolskih imenica u hrvatskome jeziku. IHJJ. Zagreb.txt	

CLOSE SAVE

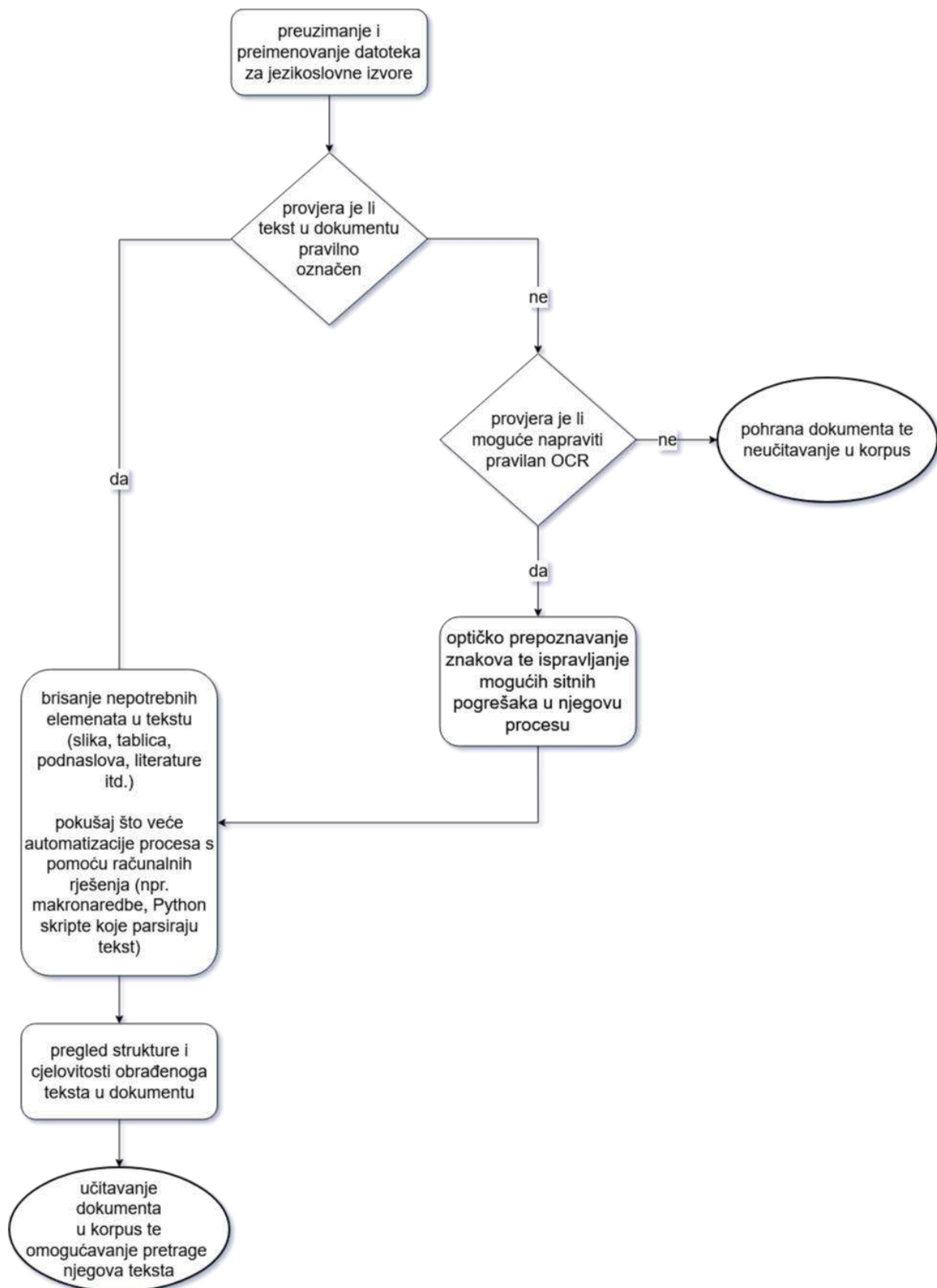
5. slika: Podatak o korpusnoj potvrdi (knjiga)

Nažalost, neki stariji brojevi časopisa sadržavaju članke koji su objavljeni kao skenirane slike bez prepoznavanja teksta, pa je bilo potrebno provesti optičko prepoznavanje znakova (OCR) s pomoću programa *ABBYY FineReader*, koji može precizno prepoznati utipkane znakove za hrvatski jezik ako se tekst u slici jasno vidi te je riječ o prepoznatljivome fontu (npr. *Times New Roman*, *Arial*, *Calibri* itd.). Optičko prepoznavanje znakova izvršilo se jedino na člancima u kojima je jasno vidljiv tekst te nema mnogo mrlja u slici i znakova koji nisu pisani latinicom (npr. tekstovi na glagoljici i ćirilici) kako bi se izbjeglo da učitani članak ima mnogo pogrešaka u tekstu. Budući da *Sketch Engine* ne može ispravno 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 formatu .pdf 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, koji imaju 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.

Točnost izrezanih članaka provjeravala se s pomoću *File Explorera* u sustavu *Windows* uključivanjem mogućnosti *Prikaz pretpregleda* (u izborniku pod *View – Preview pane*), koja omogućuje pregled dokumenta koji se ne mora otvoriti u programu za dokumenata u formatu .pdf. Svaki izrezani članak pregledan je odmah nakon izrezivanja kako bi se provjerilo da nisu dijelovi teksta u članku slučajno izrezani. Nakon toga svaki se dokument iz formata .pdf prebacio u *Word* (.docx) s pomoću *ABBYY FineReader*a, koji u procesu prebacivanja ima mogućnost da prepozna bilješke, tablice i slike 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 provela na više dokumenata. Dokumenti u *Wordu* dodatno su očišćeni te prebačeni u .txt dokumente s pomoću skripte za Python (Pythonove skripte za pročišćavanje teksta i prebacivanje u format .txt:

<https://borna12.gitlab.io/igre-mreznik/kod%20za%20jenu/pretpvarac/word%20to%20txt.py>, pristupljeno 2. siječnja 2020.), koja je u dokumentima nad kojima je provedeno optičko prepoznavanje znakova ispravila nekoliko nepoznatih znakova u tekstu (npr. ❖ u *ü*), 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 dokument u formatu *.txt* 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. Na kraju je svaki dokument u formatu *.txt* u kojemu nisu bile uočene pogreške bio učitao u korpus unutar programa *Sketch Engine*. Sažeti prikaz postupka izrade korpusa *Jene* prikazan je na 1. grafu.



1. graf: Koraci pri izradi *Jezikoslovnoga korpusa*

3. Mogućnosti pretraživanja korpusa

Korpus je još uvijek u demoinačici te još ima nekih uočljivih pogrešaka u tekstu, ali je i u postojećemu obliku veoma korisno pomagalo za sve koji se bave istraživanjem hrvatskoga jezikoslovnog nazivlja ili samo žele provjeriti tko je u svojem radu spomenuo neki naziv. Na 6. slici nalazi se prikaz konkordancije leme *poslijelog*.

The screenshot shows the 'CONCORDANCE' search interface. At the top, there is a search bar with 'Jezikoslovni' and a search icon. Below the search bar, there are navigation icons and a 'KWIC' dropdown menu. The main area displays a list of search results with columns for document ID, text, and the word 'poslijelog' in context. The results are numbered 1 through 4.

Document ID	Text	Word
1	FLUMINENSIA 29... adpozicija kojim se operira u knjizi obuhvaćaju se prijedlozi i	poslijelozni
2	FLUMINENSIA 30... 31) </s><s> Adpozicija nasuprot može se upotrebljavati kao	poslijelog
3	Rasprave 39-2, ... u takvu položaju nazvali su posijelozima, a Silić i Pranjeković	poslijelozima
4	Rasprave 39-2, ... pter nequitias spirituales razlog je odstupanja od predloška	poslijelog

6. slika: Konkordancija leme *poslijelog*

Na 7. slici nalazi se prikaz dijela skica riječi za lemu *nastavak*.

The screenshot shows the 'WORD SKETCH' interface. At the top, there is a search bar with 'Jezikoslovni' and a search icon. Below the search bar, there are navigation icons and a 'nastavak as noun 3,751x' label. The main area displays a grid of word sketches for the word 'nastavak'. Each sketch is a table with a title and a list of related words and their descriptions.

Sketch Title	Related Words and Descriptions
kakav?	padežni (padežnim nastavkom), prezentski (prezentski nastavak), genitivni (genitivni nastavak), nulti (nulti nastavak), obličan (obličnim nastavcima), infinitivan (infinitivni nastavak), starojezični (a) Starojezični nastavak), star (stari nastavak), tvoriti
subjekt_od	obilježavati (Nastavak / o / obilježava), počinjati (a prezentski nastavak počinje samoglasnikom - e), dodavati (u kojima se nastavak ka ne dodaje na čitav muški), javljati (nastavak javlja), podudarati (nastavci uglavnom podudaraju s onima iz), potvrđivati (nastavak potvrđuje), tvoriti (jednom zasvjedočeni dativni nastavak - am , tvore instrumental imenica e-sklonidbe)
koordinacija	osnova (osnove i nastavka), sufiks (sufiks i nastavak), završetak (a) imeničkim završecima ili nastavcima , b), nepromjenjiv (n u nastavcima i nepromjenjivim riječima), tvoriti (kod starodubrovačkih pisaca tvorio i nastavkom - v), nastavak (uz zadržavanje nepalatalnih nastavaka ili nastavaka ugaslih sporednih deklinacija), prijedlog (nastavcima te prijedlozima)
particip	zapisati (zapisan tipičnom kraticom SC + nastavak l), poopćiti (U imenica e-vrste poopćeni su stari palatalni nastavci u G te), zasvjedočiti (nastavak - mi zasvjedočen), ujednačiti (nastavci ujednačeni), ovjeriti (nastavak - om nije ovjeren), naglasiti (naglašen nastavak), dodati (dodan nastavak), uobičajiti (u ostalim licima uobičajeni nastavci)

7. slika: Skice riječi za lemu *nastavak*

Na 8. slici nalazi se dio popisa imenica iz korpusa.

noun

(35,913 items | 2,527,721 total frequency)

Lemma	↓	Frequency ?	Lemma	↓	Frequency ?	Lemma	↓	Frequency ?	Lemma	↓	Frequency ?
1 jezik		46,699 ...	11 govor		12,206 ...	21 skupina		9,133 ...	31 razlika		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 stojeće		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 tip		7,616 ...	36 izraz		6,497 ...
7 primjer		17,185 ...	17 rad		10,522 ...	27 knjiga		7,564 ...	37 vrsta		6,437 ...
8 imenica		16,956 ...	18 godina		9,941 ...	28 autor		7,261 ...	38 opis		6,405 ...
9 dio		15,943 ...	19 mjesto		9,546 ...	29 pitanje		7,023 ...	39 analiza		6,167 ...
10 rečenica		14,258 ...	20 pridjev		9,391 ...	30 obzir		6,960 ...	40 struktura		6,166 ...

8. slika: Dio popisa imenica iz *Jezikoslovnoga korpusa*

Na 9. slici nalazi se dio popisa naziva (ključnih riječi) automatski izlučenih iz *Jezikoslovnoga korpusa*. Nazivi se dobivaju usporedbom učestalosti pojave višerječnih skupina u *Jezikoslovnome korpusu* i u općemu korpusu *hrWaC*. Pretpostavlja se da su jezikoslovni nazivi sveze koje se češće pojavljuju u *Jezikoslovnome korpusu* nego u *hrWaC-u*.

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

KEYWORDS

Jezikoslovni



SINGLE-WORDS ✓

MULTI-WORDS ✓



reference corpus: Croatian Web (hrWaC 2.2, RFTagger)

Word	Word	Word	Word
1 imenska riječ ...	11 hrvatska književnost ...	21 hrvatski 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 nebrojiv oblik ...	33 slavenski jezik ...
4 vrsta riječi ...	14 rečenica tipa ...	24 lažan prijatelj ...	34 mjesni govor ...
5 osobno ime ...	15 srednji rod ...	25 priložna oznaka ...	35 oznaka kategorije ...
6 gledište kategorije ...	16 njemački jezik ...	26 kategorija lica i broja ...	36 značenje riječi ...
7 gledište kategorije broja ...	17 oblik sadržaja ...	27 vlastito ime ...	37 hrvatska riječ ...
8 gramatička kategorija ...	18 obiteljski nadimak ...	28 isto značenje ...	38 gramatika hrvatskoga jezika ...
9 kategorija lica ...	19 veći broj ...	29 imenska skupina ...	39 sljedeći primjer ...
10 standardan jezik ...	20 tvorba riječi ...	30 gramatička kategorija broja ...	40 ključna riječ ...

Rows per page: 50 1-50 of 1,000 1 < >

9. slika: Dio popisa ključnih riječi iz Jezikoslovnoga korpusa

Na 10. slici prikazana je usporedba uporabe pridjeva *imenički* (929 potvrda) i *imenični* (145 potvrda).

WORD SKETCH DIFFERENCE

Jezikoslovni



Get more

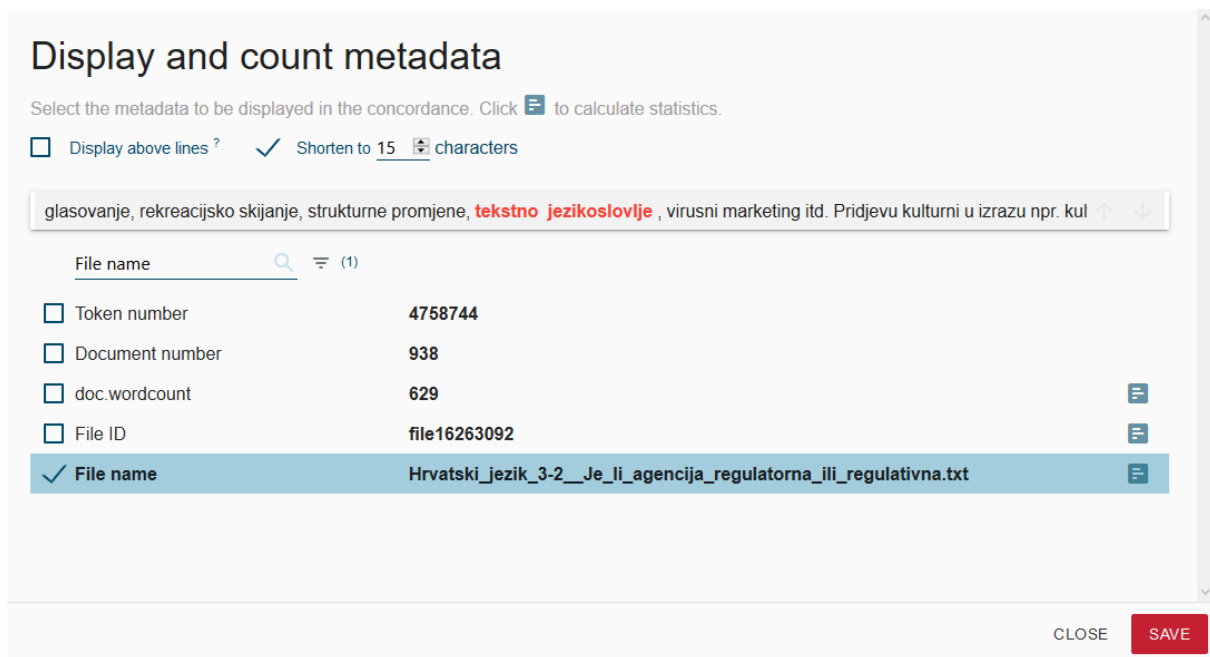
imenički 929x | imenični 145x

tko-što?	oba_u_genitivu	kako-kada?
deklinacija 27 0 ...	složenica 18 0 ...	1103 1 0 ...
klasa 25 0 ...	klasa 10 0 ...	naspram 1 0 ...
složenica 51 2 ...	referent 10 0 ...	inherentno 1 0 ...
osnova 82 7 ...	osnova 20 2 ...	polaziti 1 0 ...
sintagma 47 8 ...	posuđenica 9 2 ...	beziznimno 1 0 ...
sklonidba 29 5 ...	sintagma 13 5 ...	sukladno 1 0 ...
sastavnica 17 20 ...	i-osan 0 1 ...	samo 6 2 ...
natuknica 7 16 ...	z-osnov 0 1 ...	uglavnom 3 1 ...
z-osnov 0 1 ...	z-osan 0 1 ...	mnogo 2 1 ...
sinskup 0 2 ...	sinskup 0 2 ...	• 2 1 ...
nadregionalizam 0 2 ...	nadregionalizam 0 2 ...	često 2 2 ...
sklonidab 0 2 ...	sklonidab 0 2 ...	primjerice 1 1 ...

10. slika: Usporedba korpusne uporabe pridjeva *imenički* i *imenični*

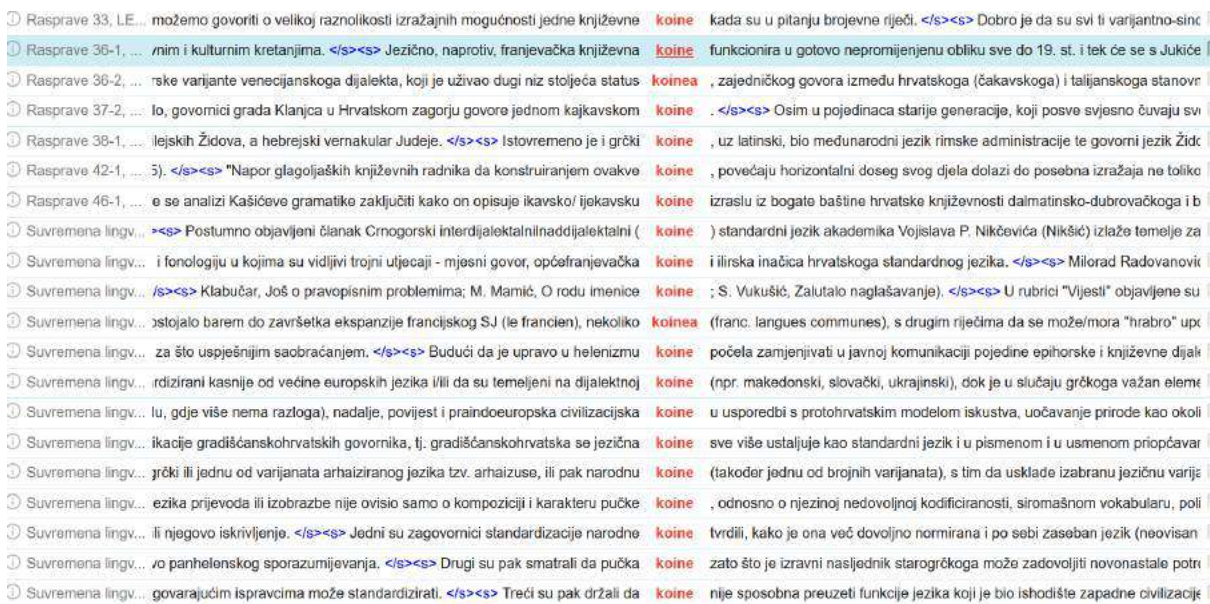
4. Zaključak

Korpus *Jene* trenutačno je još u demoinaciji te se i dalje treba dopunjavati novim tekstovima koje treba tehnički obraditi za korpus na isti način koji je prikazan u potpoglavlju *Postupak izrade korpusa*. Također, u korpusu se stalno ispravljaju uočene pogreške u prikazu teksta (npr. nedostatak razmaka nakon točke, neprepoznatljivi znakovi). Neki članici u ovoj monografiji temelje se na korpusnoj analizi *Jezikoslovnoga korpusa*. Budući da korpusu na zahtjev slanjem e-poruke Institutu za hrvatski jezik i jezikoslovlje mogu pristupiti i ostali članovi akademske zajednice koji imaju pristup programu *Sketch Engine* te korisnički račun AAI@EduHr, korpus može pomoći svima koji se bave proučavanjem jezikoslovnoga nazivlja jer se osim kolokacija (*Word Sketches*) i veza među nazivima (*Word Sketch Difference*) mogu brzo pronaći i izvori koji se bave određenim jezikoslovnim temama ili radovi u kojima se upotrebljava određeni naziv. Na primjer, na pitanje je li potvrđen naziv *tekstno jezikoslovlje*, možemo brzo odgovoriti s pomoću pronađene korpusne potvrde prikazane na 11. slici.



11. slika: Korpusna potvrda za *tekstno jezikoslovlje*

Pri razmišljanju o rodu i sklonjivosti imenice *koine* korpus je također bio nezaobilazni alat (vidi 12. sliku).



12. slika: Konkordancija imenice *koine*

Literatura

Hrvatski jezični korpus – Croatian Language Corpus. 2013. *Hrvatska jezična riznica*. <http://riznica.ihjj.hr/index.hr.html> (pristupljeno 2. rujna 2020.).

hrWaC – *Croatian web corpus*. 2013. Natural Language Processing group. <http://nlp.ffzg.hr/resources/corpora/hrwac> (pristupljeno 2. rujna 2020.).

Kilgarriff, Adam i dr. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*. Ur. Williams, Geoffrey; Vessier, Sandra. Université de Bretagne-Sud. Lorient. 105–116.

Marković, Mario; Mihaljević, Josip; Mihaljević, Milica. 2020. Kako pronaći jezikoslovni naziv. *Hrvatski jezik* 7/1. 18–22. <https://hrcak.srce.hr/235344> (pristupljeno 2. rujna 2020.).

Word sketch – collocations and word combinations. 2019. Sketch Engine Guide. <https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/> (pristupljeno 2. rujna 2020.).