

Baza podataka – Što? Kako? Zašto?

ŠTO JE BAZA PODATAKA?	2
IZGLED I PODATCI U BAZI PODATAKA	4
KRATICE VRSTA RIJEČI	5
KAKO SE SLUŽITI BAZOM PODATAKA?	6
ZAŠTO JE BAZA PODATAKA KORISNA?	9

Što je Baza podataka?

Baza podataka – Frekvencijski rječnik udžbenika prvi je rezultat projekta RAPID namijenjen široj znanstvenoj i stručnoj javnosti. Frekvencijski rječnik nastao je korpusnom analizom disciplinarnih udžbenika. Analizirana su 62 udžbenika (ukupno 91 knjiga) za predmete Matematika (1. – 6. r. OŠ), Priroda i društvo (1. – 4. r. OŠ), Priroda (5. i 6. R. OŠ) i Povijest (5. i 6. R. OŠ) triju izdavača. Broj udžbenika po predmetu i razredu prikazan je u Tablici 1.

Tablica 1 *Broj udžbenika koji čine građu korpusa RAPID prema razredu i predmetu*

	Broj knjiga (svezaka udžbenika)				
	Matematika	Priroda i društvo	Priroda	Povijest	UKUPNO
1. razred	6 (10)	5 (6)			11 (16)
2. razred	6 (10)	5 (7)			11 (17)
3. razred	6 (10)	5 (7)			11 (17)
4. razred	6 (10)	5 (7)			11 (17)
5. razred	3 (6)		3 (3)	3 (3)	9 (12)
6. razred	3 (6)		3 (3)	3 (3)	9 (12)
Ukupno	30 (52)	20 (27)	6 (6)	6 (6)	62 (91)

Kao alat za izradu korpusa odabiran je *Sketch Engine*. To je specijalizirani alat za korpusnu lingvistiku koji podržava niz jezika, uključujući i hrvatski te izradu samostalnih, korisničkih korpusa. Korpus RAPID obuhvaća ukupno 2.603.410 pojavnica (riječi koje se u tekstu pojavljuju jedna iza druge), od toga je 1.644.094 pojavnica koje su riječi, dok ostale obuhvaćaju različite znakove (npr. pravopisne), brojeve i simbole. Ukupno korpus RAPID sastoji se od 139.591 rečenice i 48.403 natuknice (temeljnoga oblika riječi).

Baza podataka – Frekvencijski rječnik sadrži popis svih obličnica (različitih gramatičkih oblika riječi) koji su pojavljuju u korpusu RAPID, a koje su prikazane prema predmetima i razredima. To znači da se pojedini oblik riječi u Bazi navodi više puta, za svaki pojedini predmet i razred u čijim se udžbenicima taj oblik riječi pojavio. Za svaki je oblik riječi u Bazi navedena oznaka vrste riječi, ukupna frekvencija (broj pojavljivanja) u udžbenicima za određeni predmet i razred. Uz to, u Bazi podataka navedene su i natuknice (citatni oblici riječi) s ukupnom frekvencijom za pojedini predmet i razred, kao i broj udžbenika (za predmet i razred) u kojima se pojavljuje. Te su informacije (uz objašnjenja) grafički prikazane na slici 1. na sljedećoj stranici.

U Bazi podataka nalaze se tzv. sirovi, odnosno neobrađeni podatci budući da je riječ o više od 350.000 obličnica koji su izravno izlučeni iz korpusa uporabom programa Sketch Engine, bez dodatne obrade. To znači da su i temeljni oblik riječi (natuknica) i vrsta riječi određeni automatski pomoću programa Sketch Engine pa je očekivan i manji broj pogrešaka, odnosno odstupanja u tim podacima.

Izgled i podatci u Bazi podataka

Oblik riječi	Riječ	Vrsta riječi	Frekvencija oblika	Frekvencija natuknice	Predmet	Razred	Zastupljenost
pisanje	pisanje	n	1	0	Matematika	1	0
	pisanje		0	1	Matematika	1	1
pisanja	pisanje	n	5	0	Matematika	2	0
pisanje	pisanje	n	13	0	Matematika	2	0
PISANJE	pisanje	n	4	0	Matematika	2	0
pisanju	pisanje	n	4	0	Matematika	2	0
	pisanje		0	26	Matematika	2	8

↓

(1) oblik riječi
(pojavnica)

↓

(2) riječ
(natuknica)

↓

(3) vrsta riječi
n = imenica

↓

(4) čestota
pojavnice

↓

(5) čestota
natuknice

↓

(6) predmet

↓

(7) razred

↓

(8) zastupljenost
u udžbenicima

Slika 1 Baza podataka

Primjer:

Oblik riječi *pisanja* (1) pripada natuknici *pisanje* (2). Taj se oblik (*pisanja*) pojavljuje ukupno 5 puta (4) u udžbenicima predmeta Matematika (6) za 2. razred osnovne škole (7). Osnovni oblik riječi jest *pisanje* (2), po vrsti riječi je imenica (3). Ta se imenica u svim svojim oblicima pojavljuje ukupno 26 puta (5) u udžbenicima Matematike (6) za 2. razred osnovne škole (7), a ukupno je zastupljena u 8 udžbenika Matematike za 2. razred (8). Oznaka 0 u tablicama znači da kategorija podatka nije dostupna. Preciznije, za oblik *pisanja* oznaka 0 nalazi se u stupcu Frekvencija natuknice i Zastupljenost. To znači da za oblik *pisanja* nije dostupan podatak o čestoti natuknice (jer to nije natuknica), a nije dostupan ni podatak o zastupljenosti u udžbenicima.

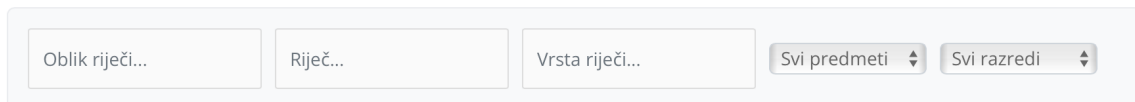
Kratice vrsta riječi

Kratice vrsta riječi preuzete su iz programa Sketch Engine, a kao što je već navedeno, tijekom obrade korpusa program samostalno određuje vrste riječi. Oznake u stupcu Vrsa riječi imaju sljedeće značenje:

Oznaka	Značenje
N	imenica
V	glagol
A	pridjev
P	zamjenica
R	prilog
S	prijedlog
C	veznik
M	broj
Q	čestica
I	uzvik
Y	kratica
X	ostalo
Z	interpunkcija

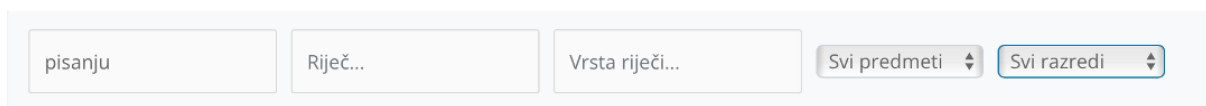
Kako se služiti Bazom podataka?

Baza podataka sadrži tražilicu koja omogućuje pretragu i filtriranje podataka po nekoliko kategorija: oblik riječi, riječ, vrsta riječi, predmet i razred (slika 2), kao i izvoz rezultata na nekoliko načina.



Slika 2 Izgled tražilice u Bazi podataka

Moguće je pretražiti bazu kako bi se saznali podatci o određenom obliku riječi. Na primjer, ako u tražilicu u polje oblik riječi upišemo *pisanju* (slika 3), dobit ćemo rezultat kao na slici 4. Slika 4 prikazuje da se oblik pisanje pojavljuje u više predmeta i više razreda, za svaki predmet i razred prikazane su frekvencije toga oblika.



Slika 3 Pretraga za oblik riječi *pisanju*

Oblik riječi	Riječ	Vrsta riječi	Frekvencija oblika	Frekvencija natuknice	Predmet	Razred	Zastupljenost
pisanju	pisanje	n	4	0	Matematika	2	0
pisanju	pisanje	n	2	0	Matematika	3	0
pisanju	pisanje	n	3	0	Matematika	4	0
pisanju	pisanje	n	1	0	Matematika	5	0
pisanju	pisanje	n	1	0	Matematika	6	0
PISANJU	pisanje	n	1	0	Priroda i društvo	1	0

Slika 4 Dio rezultata pretrage oblika riječi *pisanju*

Pomoću tražilice moguće je pretražiti podatke za temeljni oblik te riječi tako što se u polje Riječ unese *pisanje*, kao na slici 5. Tada se uz obličnicu *pisanje*, pojavljuje i natuknica *pisanje* uz podatke o zastupljenosti natuknice, kao što je vidljivo na slici 6.

Oblik riječi...

pisanje

Vrsta riječi...

Svi predmeti ▾

Svi razredi ▾

Slika 5 Pretraga riječi *pisanje*

	pisanje		0	54	Matematika	3	6
pisanja	pisanje	n	2	0	Matematika	4	0
pisanje	pisanje	n	14	0	Matematika	4	0
PISANJE	pisanje	n	12	0	Matematika	4	0
pisanju	pisanje	n	3	0	Matematika	4	0
	pisanje		0	31	Matematika	4	9

Slika 6 Rezultati pretrage za riječ *pisanje*

Pretragu je moguće ograničiti s obzirom na razred. Primjerice, slika 7 prikazuje način pretrage, a slika 8 rezultate zastupljenosti oblika riječi *pisanju* u udžbenicima za 4. razred.

pisanju

Riječ...

Vrsta riječi...

Svi predmeti ▾

4. razred ▾

Slika 7 Pretraga oblika riječi *pisanju* u udžbenicima za 4. razred

Oblik riječi	Riječ	Vrsta riječi	Frekvencija oblika	Frekvencija natuknice	Predmet	Razred	Zastupljenost
pisanju	pisanje	n	3	0	Matematika	4	0

Slika 8 Zastupljenost oblika *pisanju* u udžbenicima za 4. razred

Pretragu je moguće ograničiti s obzirom na predmet. Primjerice, slika 9 prikazuje rezultate zastupljenosti oblika riječi *pisanju* u udžbenicima za Povijest.

Slika 9 Pretraga oblika riječi *pisanju* u udžbenicima za predmet Povijest

Oblik riječi	Riječ	Vrsta riječi	Frekvencija oblika	Frekvencija natuknice	Predmet	Razred	Zastupljenost
PISANJU	pisanje	n	1	0	Priroda i društvo	1	0
pisanju	pisanje	n	2	0	Priroda i društvo	2	0

Slika 10 Zastupljenost oblika riječi *pisanju* u udžbenicima za predmet Povijest

Podatke dobivene pretragom Baze moguće je izvesti na nekoliko načina, kao što je prikazano na Slici 11. Podatke je moguće kopirati, izvesti kao Excell dokument, kao PdF ili CSV. Također ih je moguće i ispisati. Korisnici također mogu odabrati koliko redaka po stranici žele prikazati.

Kopiraj
CSV
Excel
PDF
Ispis

Prikaži redaka

Pretraži:

Slika 11 Načini izvoza rezultata pretrage

Zašto je Baza podataka korisna?

Izrađena Baza podataka može služiti u istraživačke svrhe ukoliko istraživače zanima leksik disciplinarnih udžbenika koji se istražuju u projektu RAPID i to na razini obličnica i natuknica. Pretragom Baze moguće je istraživati zastupljenost pojedinih riječi u razredima i nastavnim predmetima, kao i to u kojim se sve oblicima pojedine riječi pojavljuju.

Ova Baza također može biti korisna i učiteljima na više načina. Pomoću Baze podataka učitelji mogu dobiti uvid u rječnik kojim učenici trebaju ovladati u pojedinim predmetima, podatci o čestoti pojedinih riječi i oblika, kao i njihovoj zastupljenosti mogu im pomoći pri planiranju nastave, posebno različitih vrsta prilagodbe određenim skupinama učenika. Također je moguće služiti se bazom i kao izvorom građe za izradu različitih jezičnih zadataka, primjerice pravopisnih.