



Konferencija Srce DEI



HRABAR: Akademski korpus bogat metapodacima za OPJ izveden iz Hrčka i Dabra

Ivan Grubišić*, Miha Keber*, Damir Korenčić, Tomislav Šmuc i Anja Barešić

Institut Ruđer Bošković

*Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva

Srce DEI 2026





Uvod

- Hrčak & Dabar → **HRABAR** paralelni sažetci hrvatsko-engleskih tekstnih podataka
- Prikupljanje .xml datoteka
- Čišćenje i standardizacija
- Razdvajanje rečenica primjenom većinskog glasovanja
- TMX poravnanje razdvojenih segmenata
- Morfosintaktičke oznake pomoću *classla* i *stanza* alata

DABAR - Digital Academic Archives and Repositories

Search the content of repository



[Advanced search](#)

 BROWSE

 UPLOAD



186

Broj repozitorija

338 810

Published objects

57.2 %

Open Access

Portal hrvatskih znanstvenih i stručnih časopisa

Pretraži

[Napredno pretraživanje](#) [Upute za pretraživanje](#)

Hrčak je centralni portal koji na jednom mjestu okuplja hrvatske znanstvene i stručne časopise koji nude otvoreni pristup svojim radovima ([više](#)).

575

Časopisa

24.878

Sveščića

323.269

Radova s cjelovitim tekstom

103.974

ORCID identifikatora



Prikupljanje podataka



- <https://dabar.srce.hr/>
- <https://hrcak.srce.hr/>

Čišćenje sažetaka, ključnih riječi, itd.

Izvučen tekst iz XML-a:

Ovaj diplomski rad pojašnjava mogućnosti povezivanja dvaju srodnih **umjetničkih** područja: glume i glazbe. U pedagoškom radu pojedinih profesora s odsjeka Glume na Akademiji dramske umjetnosti **Sveučilišta** u Zagrebu već postoje pokušaji korelacije i primjene znanja iz glazbe u **glumačkom** procesu i obrnuto,

{ => šć

} => ć

{ => š

~ => č

è => č

aea => ća

aeu => ću

aei => ći

aeë => će

aeo => ćo

Procesirani tekst:

Ovaj diplomski rad pojašnjava mogućnosti povezivanja dvaju srodnih **umjetničkih** područja: glume i glazbe. U pedagoškom radu pojedinih profesora s odsjeka Glume na Akademiji dramske umjetnosti **Sveučilišta** u Zagrebu već postoje pokušaji korelacije i primjene znanja iz glazbe u **glumačkom** procesu i obrnuto,

Čišćenje Dabar korpusa	Ukupno sažetaka	uklonjeno	Promijenjeno HR	Promijenjeno EN
original	299.387	-	-	-
remove entries with repeating chars	299.387	-	280.136	91.197
clean by absolute length	299.387	-	281.086	91.672
add missing abstracts	299.387	-	281.086	91.672
clean by absolute length	299.387	-	281.086	91.672
drop none duplicate	259.657	39.730	-	-
clean br	259.657	-	-	-
clean hyphens	259.657	-	-	-
clean rnt	259.657	-	-	-
clean words	259.657	-	7.141	1.779
clean sc	259.657	-	209	303
correct words	259.657	-	151	41
clean entries with repeating texts	259.657	-	88	74
clean by absolute length	259.657	-	6	11
drop none duplicate	259.408	249	-	-
remove by relative length	257.650	1.758	-	-
clean language	257.650	-	3.385	62.122
drop none duplicate	192.897	64.753	-	-
Ukupno	192.897	106.490	1.066.906	547.329

vodaei => vodeći



Usporedba veličine skupa podataka



Dabar

	sažeci	znakovi	tokeni	rečenice
hrvatski tekstovi	262.287	330.339.379	51.472.049	2.136.652
engleski tekstovi	198.079	269.422.523	46.593.085	1.701.453



Hrčak

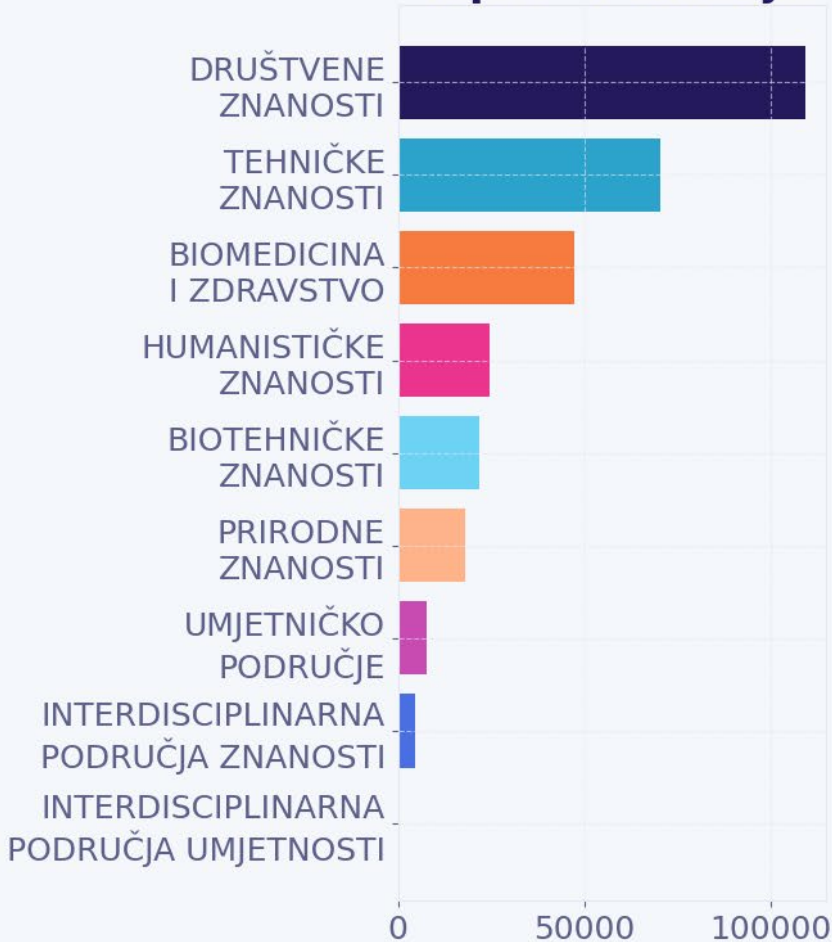
	sažeci	znakovi	tokeni	rečenice
hrvatski tekstovi	120.055	126.813.510	20.091.304	762.739
engleski tekstovi	94.110	121.352.422	21.284.331	713.528



nvzz.hr - Dabar

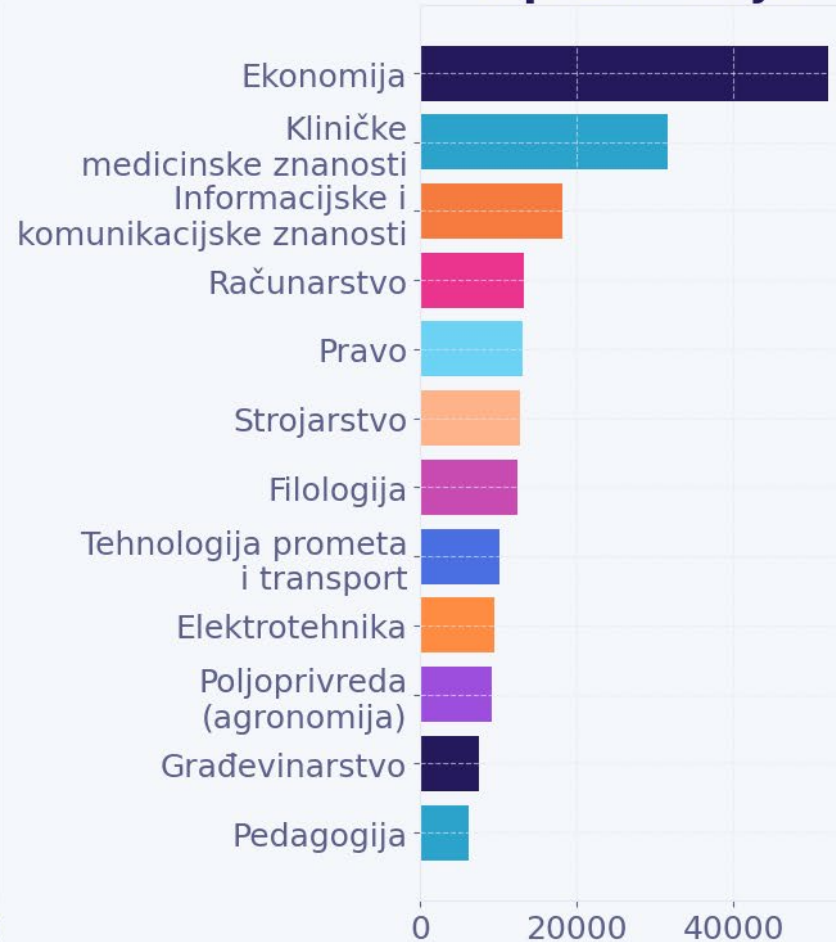
Ukupno: 304.593

Top 9 - Područje



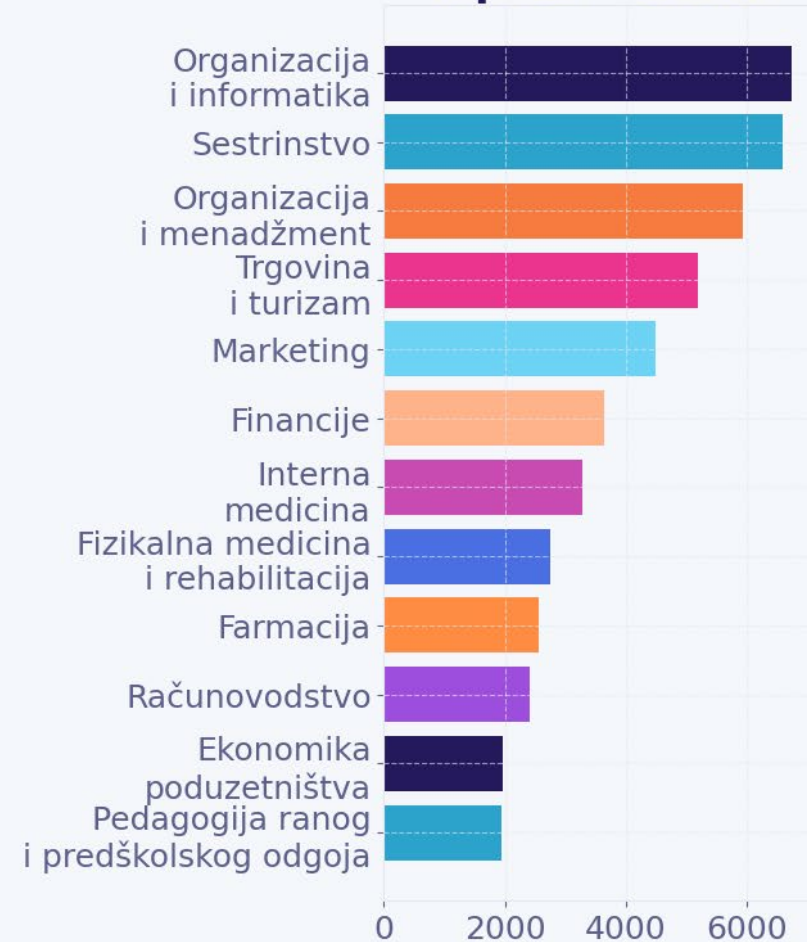
Ukupno: 304.018

Top 12 - Polje



Ukupno: 153.872

Top 12 - Grana





Razdvajanje rečenica

- Razdvajanje s glasovanjem ansambla postojećih modela
 - Classla, stanza, spacy
- Odaberemo glasanje svih modela
- Razdvojene rečenice → paralelne segmente teksta

1.6%
log
skala





Poravnanje segmenata teksta

- SentAlign – semantička sličnost pomoću „Language-agnostic BERT Sentence Embedding” [1],
- koristi heuristiku za optimizaciju alignmenta [2],
- Ukupan broj poravnatih segmenata teksta:
 - **Dabar TUs: 1.540.520**
 - **Hrčak TUs: 1.117.034**

[1] Feng, Fangxiaoyu, et al. "Language-agnostic BERT sentence embedding." Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers). 2022.

[2] Steingrímsson, Steinþór, Hrafn Loftsson, and Andy Way. "SentAlign: Accurate and scalable sentence alignment." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2023.



Primjena

- Razvoj i testiranje modela i **metoda za čišćenje podataka**
- Podatci za **hrvatsko-engleski strojni prijevod**
- Hijerarhijska klasifikacija nvzz.hr: **Područje>Polje>Grana**
- Hrvatsko-engleske **ključne riječi**
- **Podatkovna znanost** na tekstnim podacima i metapodacima.



Autori i suradnici s HRABAR-a:

- Ivan Grubišić
- Dr. sc. Anja Barešić
- Dr. sc. Tomislav Šmuc
- Dr. sc. Damir Korenčić

Mentori:

dr. sc. Anja Barešić

i

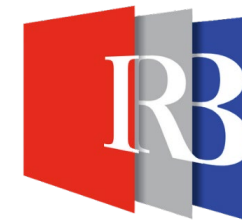
prof. dr. sc. Alan Jović



Ovo djelo je dano na korištenje pod licencom Creative Commons
Imenovanje 4.0 međunarodna.

www.srce.unizg.hr

creativecommons.org/licenses/by/4.0/deed



- SRCE,  Hrčak i  Dabar
- Projekt:



Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

www.srce.unizg.hr/otvoreni-pristup

