



arnes 



**NI4OS**  
National Initiatives for Open Science in Europe

# Jezikoslovna infrastruktura za slovenščino: CLARIn.SI in RI-SI - CLARIN

Darinka Verdonik, UM FERl

Mreža znanja 2021, od 23. do 30. novembra



CLARIN.SI je slovenski nacionalni konzorcij v mreži evropske raziskovalne infrastrukture [CLARIN](#).

Raziskovalcem na področju humanistike, družboslovja in drugih, z jezikom povezanih ved zagotavlja:

- jezikovne vire in tehnologije ter
- strokovno podporo in prenos znanja.

Na ta način želi razširiti tehnološke možnosti raziskovanja jezikov, predvsem slovenščine in drugih južnoslovanskih jezikov, ter spodbujati meddisciplinarno sodelovanje.

# Računalniško podprto raziskovanje jezika: jezikoslovje in tehniške vede

## KORPUSI

- referenčni
- specializirani
- prevodi

## GOVORNE BAZE

- posnetki govora
- zapis govora

## PODATKI O BESEDIŠČU

- sezname besed, sopomenk, kolokacij, koligacij...

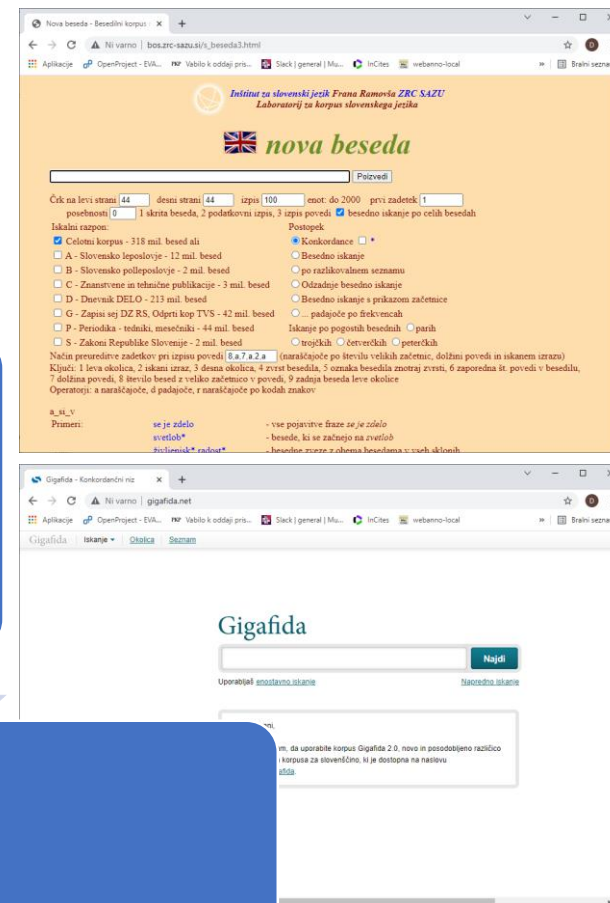
## ORODJA

- tokenizator in lematizator
- oblikoslovni označevalnik
- skladenjski razčlenjevalnik
- označevanje imenskih entitet
- semantična analiza...

# Zgodovinski pregled

Stihijski razvoj -> istovrstni vir re je razvijal vedno znova

Sodelovanje, prostodostopna in odprta infrastruktura





# The research infrastructure for language as social and cultural data

CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.



# CLARIN – ERIC



Find

# Linguistic Data and NLP Tools

Citation Support (with Persistent IDs)

CLARIN.SI



Išči

[Napredno iskanje](#)

### Avtor

Ljubešič, Nikola (95)

Erjavec, Tomaž (71)

Krek, Simon (41)

Arhar Holdt, Špela (37)

Dobrovoljc, Kaja (34)

... pogledajte več

### Ključna beseda

TEI (44)

language model (33)

manual annotation (33)

computer-mediated co ... (30)

lemmatisation (29)

... pogledajte več

### Jezik (ISO)

Slovenian (170)

Croatian (42)

English (37)

Serbian (27)

Bulgarian (15)

... pogledajte več

# CLARIN.SI

### What's New

Corpus

CLARIN.SI Data & Tools

Ta stran uporablja piškotke. Z nadaljevanjem uporabe te strani soglašate z uporabo piškotkov.

Razumen

# REPOZITORIJ CLARIN.SI

The screenshot shows the NoSketch Engine website. The header includes the CLARIN.SI logo and the text "NoSketch Engine" with a language selector set to "English". Below the header, there is a navigation menu with links for "CLARIN.SI / NoSketch Engine", "Raziskovalna infrastruktura CLARIN.SI", and "KonText". A search bar is located on the left side. Below the search bar, there is a list of search filters: "Vsli korpusi", "slovenski:", "referenčni", "specializirani", "spletni", "vzporedni", "parlamentarni", "ParlaMint", and "japonski". At the bottom, there is a "Preglednica" (Table) with columns for "išči", "vir", "jezik", and "zvrsta".

išči	vir	jezik	zvrsta
metaFida v0.1 (združeni korpus)	Info	Slovenian	C/A
GigaFida v2.0 proto (referenčni, nededupliciran)	Info	Slovenian	C/A

The screenshot shows the KonText website. The header includes the CLARIN.SI logo and the text "Repository About Contact" with a language selector set to "Slovene". Below the header, there is a navigation menu with links for "kon text", "Poizvedba", "Korpusi", "Shrani", "Konkordance", "Filter", "Frekvenca", "Kolokacije", "View", and "Pomoč". Below the navigation menu, there is a search bar and a list of search filters: "Slovene corpora", "Other languages", and "Parallel corpora". At the bottom, there is a "Show by size" button.

The screenshot shows the CLARIN.SI repository information page. The header includes the CLARIN.SI logo and the text "Slovenska raziskovalna infrastruktura za jezikovne vire in tehnologije" and "Common Language Resources and Technology Infrastructure, Slovenia". Below the header, there is a section titled "O REPOZITORIJU CLARIN.SI" with a description of the repository. To the right, there is a section titled "OBVESTILA:" with a notice about the "Nagrada Stevena Krauerja".

## O REPOZITORIJU CLARIN.SI

Eden izmed osnovnih namenov infrastrukture CLARIN je zagotavljanje zanesljivega arhiviranja in dostopa do jezikovnih virov, kot so korpusi, leksikoni, avdio- in videoposnetki, slovnice, jezikovni modeli itd.

CLARIN.SI vzdržuje certificiran repozitorij, v katerem je trenutno deponiranih prek 200 jezikovnih virov in orodij oz. približno 200 GB podatkov za 80 jezikov, pri čemer je večinski del namenjen slovenščini, ki ji sledita hrvaščina in srbščina. Repozitorij vsebuje širok nabor večjih korpusov (tj. urejenih zbirk besedil), primernih za raziskovanje slovenščine, pa tudi več vzporednih in ročno označenih korpusov ter leksikonov in jezikovnih modelov za uporabo v jezikovnih orodjih.

### OBVESTILA:

#### Nagrada Stevena Krauerja

Z veseljem sporočamo, da je Tomaž Erjavec dobitnik letošnje nagrade [Steven Krauer](#) za izjemne prispevke k ciljem infrastrukture CLARIN ERIC. Nagrada je bila podeljena v okvirju letne konference CLARIN (CLARIN Annual Conference 2021). Več informacij o nagradi in njenem dobitniku je dostopno na tej [povezavi](#).

# RI-SI - CLARIN

## CILJI

- zagotoviti nadaljnje delovanje tehničnih storitev infrastrukture CLARIN.SI
- omogočiti hranjenje velikih multimodalnih jezikovnih podatkov
- omogočiti, da CLARIN.SI sledi paradigmi »velepodatkov« (angl. big data)
- omogočiti, da CLARIN.SI ponuja javno dostopne spletne storitve za obdelavo velikih količin slovenskih besedil
- vzpostaviti namensko gručo računalnikov s pospeševalniki GPGPU za potrebe globokega učenja



# NABAVLJENA OPREMA

## INSITUT JOŽEF STEFAN

- Gruča za spletne storitve
- Strežnik repozitorija
- Diskovno polje za varnostne kopije
- Stikalo za optični kanal

## UNIVERZA V MARIBORU

- NVIDIA DGX-1
- Strežniška infrastruktura za obdelavo in hranjenje velepodatkovnih jezikovnih virov

## UNIVERZA V LJUBLJANI

- Strežniška rezina

# ZAKLJUČEK

- Predstavili smo odprto in prosto dostopno jezikoslovno infrastrukturo CLARIN.SI.
- V jezikoslovju in jezikovnih tehnologijah je bilo odpiranje podatkov in oblikovanje skupnega konzorcija temeljni korak k učinkovitejšemu in ekonomičnejšemu razvoju področja.

Hvala za pozornost  
darinka.verdonik@um.si

arnes 