

# Pregled storitev CLARIN.SI za ravnanje z jezikovnimi raziskovalnimi podatki

Kaja Dobrovoljc  
Filozofska fakulteta UL  
Institut Jožef Stefan

Letno srečanje Doktorske šole UL 2022  
UL FDV, 2. junij 2022



# Nekaj primerov jezikovnih podatkov

## Korpus pisnih besedil Gigafida

- časopisi, revije, spletne novice, priročniki, leposlovje
- 38.000 besedil
- 1 milijarda besed



## Korpus starejših besedil IMP

- besedila od konca 16. stol. do 1918
- 650 enot
- 14 milijonov besed



## Korpus spletnih besedil JANES

- tviti, besedila forumov, komentarji na novice, blogi ...
- 13 milijonov zapisov, 252 milijonov besed



## Korpus parlamentarnih razprav ParlaMint

- parlamentarne razprave v 17 jezikih
- za slovenščino 2014-2022
- 23 milijonov besed

ParlaMint



# UPORABA OBSTOJEČIH PODATKOV

# Pregled obstoječih podatkovnih zbirk

- zbiranje, obdelava in objava novih podatkov vzamejo **veliko časa** in zahtevajo **dodatna znanja**
- uporabni vstopni točki za pregled obstoječih jezikovnih raziskovalnih podatkov:
  - globalno: **Virtual Language Observatory**
  - lokalno: **repozitorij CLARIN.SI**

If I have seen  
further it is by  
standing on the  
shoulders of Giants



# Iskanje po repozitoriju CLARIN.SI

- <https://www.clarin.si/repository/xmlui/>
- stalna in varna hramba jezikovnih virov (tudi) za slovenščino; več kot 300 vnosov

Napredno iskanje

Išči

Omejite svoje iskanje

Avtor

Ključna beseda

Pravice

Jezik (ISO)

Vrsta

text (292)

lexicalConceptualResource (153)

corpus (147)

toolService (44)

audio (8)

languageDescription (2)

image (1)

Prikazovanje 1–10 od 346 zadetkov

1 2 3 > 35

Corpus

CLARIN.SI Data & Tools

Slovene web corpus MaCoCu-sl 1.0

(Jožef Stefan Institute; Prompsit; Rijksuniversiteit Groningen; Universitat d'Alacant / 2022-04-29)

Avtorji:

Bañón, Marta ; et al.

► prikaži vse

Ta vnos vsebuje 3 datotek(e) (12.9 GB).

Corpus

Croatian web corpus MaCoCu-hr 1.0

(Jožef Stefan Institute; Prompsit; Rijksuniversiteit Groningen;

CLARIN.SI

Brskanje

> Celoten repozitorij

Moj račun

Prijava

Splošne informacije

O vnosu v repozitorij

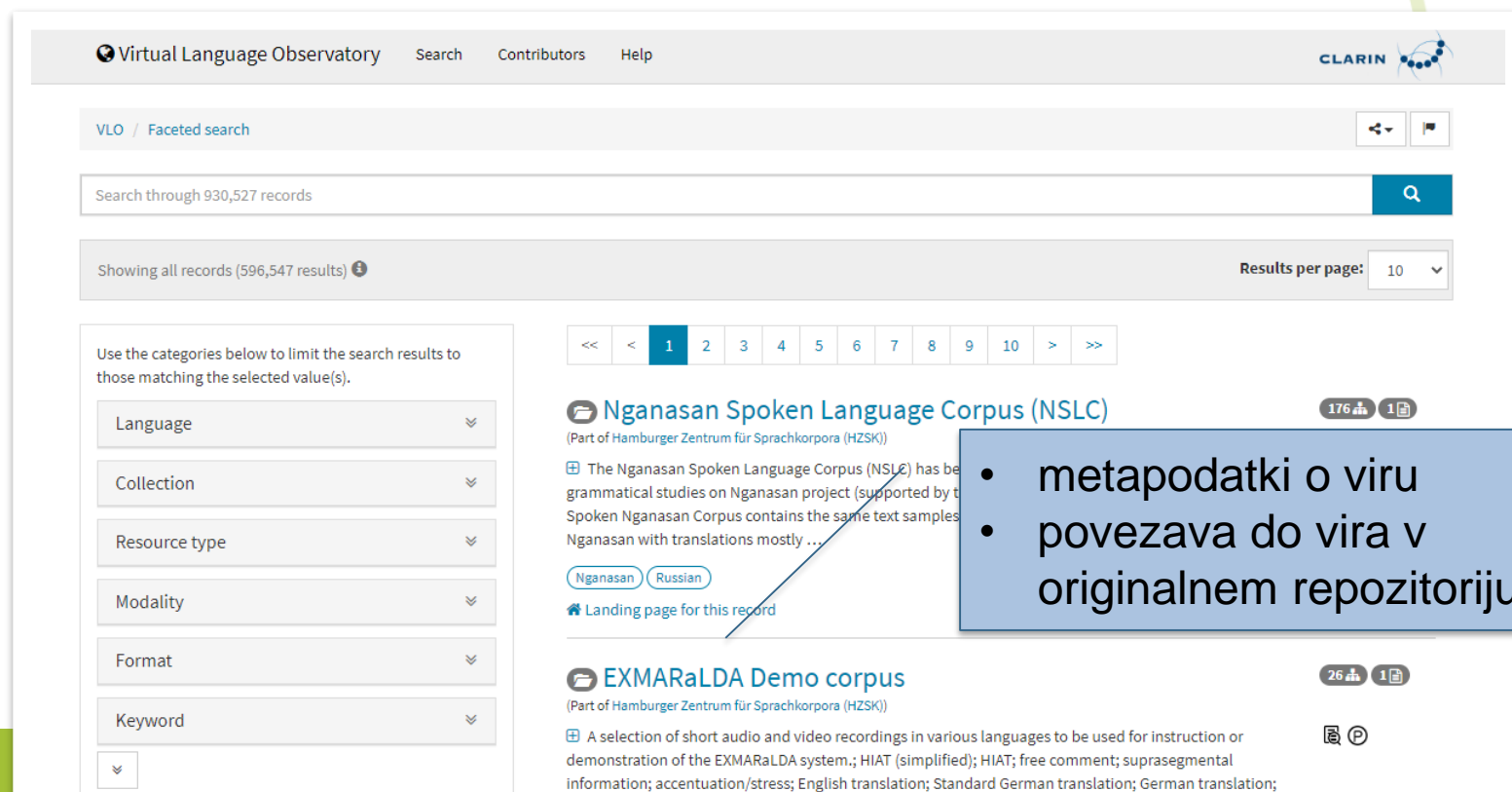
Citiranje

Življenjski cikel vnosa

Iskanje po ključnih besedah in različnih vrstah metapodatkov.

# Iskanje po Virtual Language Observatory

- [vlo.clarin.eu](http://vlo.clarin.eu)
- skupni iskalnik po vseh jezikovnih virih v nacionalnih repozitorijih CLARIN in drugih podatkovnih repozitorijih



Virtual Language Observatory Search Contributors Help CLARIN

VLO / Faceted search

Search through 930,527 records

Showing all records (596,547 results) Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Collection

Resource type

Modality

Format

Keyword

<< < 1 2 3 4 5 6 7 8 9 10 > >>

**Nganasan Spoken Language Corpus (NSLC)** 176 1

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

The Nganasan Spoken Language Corpus (NSLC) has been used for grammatical studies on Nganasan project (supported by the German Research Foundation). The Spoken Nganasan Corpus contains the same text samples as the written Nganasan with translations mostly in German.

Nganasan Russian

Landing page for this record

**EXMARaLDA Demo corpus** 26 1

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

A selection of short audio and video recordings in various languages to be used for instruction or demonstration of the EXMARaLDA system.; HIAT (simplified); HIAT; free comment; suprasegmental information; accentuation/stress; English translation; Standard German translation; German translation;

- metapodatki o viru
- povezava do vira v originalnem repozitoriju

# Brskanje po korpusih CLARIN.SI

- poleg prevzema datotek običajno v opisu vira navedene tudi **druge možnosti dostopa** do vsebovanih podatkov (npr. spletni vmesniki, priporočena orodja)
- za **iskanje po besedilnih korpusih** podporo nudi tudi CLARIN.SI preko dveh spletnih konkordančnikov:
  - noSketchEngine: [clarin.si/noske](http://clarin.si/noske)
  - Kontext: [clarin.si/kontext](http://clarin.si/kontext)

# Primer iskanja po noSketchEngine

The screenshot shows the noSketchEngine search interface. The search bar contains the word "lahko" and the engine version is "Gos 1.1.1 (referenčni, govorni)". The search results are displayed on page 1 of 209. The results list various contexts where the word "lahko" is used, such as "ja no no tako ja ja ja ja bom ja... mhm ... m boš lahko [ime] ti tudi še malo za podatke... eee torej" and "odzuni sedim mislm eem eem Ementalerja jah loh ti gam prnesu če boš dol prletu eem je pa unle bom". The interface includes a sidebar with navigation options like Home, Search, Word list, Corpus info, My jobs, User guide, Save, Make subcorpus, View options, Sort, Filter, and Frequency. A blue box highlights the search results and the filter options.

Query **lahko** 4,180 > Shuffle 4,180 > Shuffle 4,180 > Shuffle 4,180 > Shuffle 4,180 > Shuffle 4,180 (3,929.08 per million) ⓘ

Page 1 of 209 Go Next Last

gos174 ja no no tako ja ja ja ja bom ja... mhm ... m boš **lahko** [ime] ti tudi še malo za podatke... eee torej  
gos281 odzuni sedim mislm eem eem Ementalerja jah **loh** ti gam prnesu če boš dol prletu eem je pa unle bom  
gos066 ble m možne ko bi mogle bit in sem pač mislo da bom **lahko** to uspel in ni šlo po domače rečeno skozi  
gos213 tu sma rekla samo glede eee težav eee to se **lahko** uporabla za hujšanje eee pridobivanje teže  
gos045 v mestu eee nič pa se ne nardi da bi se kolesarji **lažje** vozli res v mestu da bi ne bi bil tak hudi promet  
gos184 ob sej itak ne priš delat dvanajstih delat **lahko** bi šla be ti pa mati bošte farbale kaj mo farbale?  
gos139 ma bite ka na rej d prijd ka trbelo kurte v pekle **leka** bov to mal zatog stavla ka tan prejk več nega  
gos072 načrtno najti strategije eee kaj je tisto kar **lahko** prinaša tisto dodatno vrednost konkurenčnost  
gos276 k d ti pišče vseke pol minute?... aha ... loh ma ja **loh** de je kr ja ja ja mism nje ma minutni kašn ma vsek  
gos181 ta gospod k organizira prav koko a sam uraden **lah** vozjo a sam uraden lah to delajo tud mi nismo ker  
gos198 kaj t eee eee tud to sprej karkoli nardimo? kaj to **lahko** eee potegnem malo narazen malo narazen ja z  
gos176 mism ampak ni ne... ne jz sem bla na stojšču ne sej **lahko** se pa zmenmo z njim pa na dvakrat jz nisl še nikol  
gos225 naslednjič bo mel on spet isti smeh zej pa spet **lah** je spet isto lah ni isto veš trikrat štirkat on  
gos001 medtem ko vojvodski prestol pa stoji a ne vsak ga **lahko** vidi ob cesti kekšnih deset kilometrov severno  
gos178 bo ne mislim to vseen pač neki ludi more bit a a si **lahka** jz te stvari kr kr pišem t vsej da si vsaj seveda  
gos188 prispevkov na žalost ampak odgovore boste **lahko** na primer v jezikovnem kotičku so študentje  
gos136 eni strani zato ker nam sodelovanje v tej odaji **lahko** pomeni dve stvari ne eno je neko tako osebno  
gos123 policiste civilne strokovnjake in druge ki bi **lahko** sodelovali v različnih mednarodnih  
gos271 pa on ja ja itak sm ful sit pa jz tk ja v rei glej se ti **lah** še nadeam ne se ni panike ja ja... .. kak je pa [ime]  
gos041 pravice načelno od primera do primera pa **lahko** tudi kekšen drug od teh primerov v zadnem času

Page 1 of 209 Go Next Last

- prikaz vseh zadetkov iskanja v kontekstu
- možnost dodatnega filtriranja
- statistični sezname
- vizualizacija statističnih podatkov
- možnost izvoza zadetkov



# ZBIRANJE NOVIH PODATKOV

# Zbiranje jezikovnih podatkov

- številne klasične in inovativne metode
  - zbiranje besedil
  - snemanje in transkripcija
  - vprašalniki
  - mobilne aplikacije
  - spletno pajkanje
  - ...

# Pravni vidiki – koristne povezave

- uporabne pravne informacije o **avtorskih pravicah, licenciranju in varstvu osebnih podatkov**:  
<https://www.clarin.eu/content/legal-information-platform>
- orodje za pomoč pri oblikovanju soglasja za sodelovanje v raziskavi (GDPR): <https://consent.dariah.eu>
- [vzorec pogodbe CLARIN.SI](#) za prenos avtorskih pravic za namene izdelave in objave korpusa

# Načini zapisa podatkov

- izbira načina zapisa podatkov (formata) je ključna za zagotavljanje njihove nadaljnje uporabnosti
  - izkoristite tehnično podporo
- nekaj [priporočil](#) CLARIN.SI:
  - strojna berljivost in pretvorljivost
  - uporaba standardov
  - dobre prakse: .xml, .csv, .tsv, .txt ...
  - odsvetovano: .docx, .xlsx, .pdf
- enako pomemben tudi zapis metapodatkov

# Hranjenje podatkov

- sprotna in sistematična **dokumentacija**
  - povedna imena datotek, eksplicitne pripone
  - številčenje različic, stiskanje (.zip)
  - beleženje sprememb v obliki dnevnika
- **varnostna kopija** podatkov
  - 1TB prostora v [oblaku OneDrive](#) za vse z identiteto UL

# Označevanje besedil (opcijsko)

- golim besedilom lahko na ravni besed, stavkov ali celotnih dokumentov strojno ali ročno pripišemo tudi različne oznake oz. kategorije
- nekaj tipičnih primerov:
  - slovnične lastnosti besed
  - sentiment stavkov
  - tema besedila
- številne prednosti označenih besedil
  - **lažje iskanje** (priklic zadetkov po oznakah)
  - **kvalitetnejša analiza** podatkov
  - računalniška uporaba (strojno učenje)

# Označevanje besedil (opcijsko)

- za avtomatsko splošno procesiranje smiselno preveriti že obstoječa računalniška orodja
  - CLARIN.SI Repozitorij > toolService
  - CLARIN.SI GitHub: <https://github.com/clarinsi>
  - CLARIN Language Resource Switchboard: <https://switchboard.clarin.eu/>
- za ročno označevanje na voljo spletno orodje WebAnno
  - <https://www.clarin.si/webanno/>

The screenshot shows the WebAnno interface. At the top, there is a red header with the word "Curation" and a "WebAnno | Home" link. Below the header is a navigation bar with various icons and labels: "Document" (Open, Re-create, Merge, Prev, Next, Export, Settings), "Page" (First, Prev, Go to 7, Next, Last), "Script" (LTR/RTL), "Help" (Guidelines), and "Workflow" (Finish). The main content area is divided into "Sentences" and "Annotation". The "Sentences" list shows sentence 1 highlighted in red. The "Annotation" area displays a sentence: "7 Avtorja pravita, da parameter Lndim zajema vpliv učinkov ostenja ( angl. boundary effects ), zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .". Above this sentence, there are two terms: "2TermSlv" and "1TermEng", connected by a red arrow labeled "(kas Translation)". The "Actions" panel on the right shows "Layer kas.BiTerm", "Forward annotation ?" with a checkbox, and "Annotation No annotation selected!". The bottom status bar indicates "Showing 7-8 of 41 sentences [document 23 of 50]".

# **OBJAVA PODATKOV**

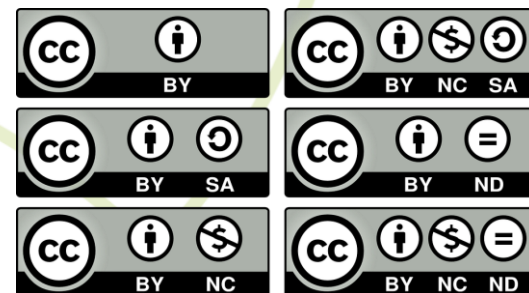


# Objava vira v repozitoriju CLARIN.SI

1. vnos ustvari avtor ([Ustvarite nov vnos](#))
  2. prijava preko ponudnika identitete (EduGain)
  3. urednik pregleda, opozori na popravke in objavi
  4. vir postane viden navzven
    - tudi preko Google, VLO, DataCite, arXive ...
    - korpusi običajno integrirani tudi v konkordančnike
- pred vnosom nujno prebrati dokumentacijo
    - [Kako ustvariti vnos](#)
    - [Pogosta vprašanja](#)

# Licence

- ob vnosu vira morajo avtorji opredeliti tudi licenco
- velika večina virov je dostopna pod eno od licenc Creative Commons (CC)
- mogoče je izbrati tudi [številne druge](#) ali ustvariti novo
- v procesu objave v pomoč **‘OPEN License Selector’**



# Razpis za projekte CLARIN.SI

- [Letni razpis CLARIN.SI](#), namenjen izdelavi ali nadgradnji virov ali storitev, ki pripomorejo k uresničevanju usmeritev infrastrukture CLARIN(.SI)
  - izdelava ali nadgradnja virov, spletnih storitev ali programske opreme
  - organizacija izobraževalnih dogodkov oz. priprava izobraževalnih gradiv
  - raziskave, ki uporabljajo obstoječe vire ali storitve CLARIN(.SI)
- rok za prijave običajno **spomladi**
- v 2022 sredstva v višini **2.000-10.000 EUR**

# Obrazec NRRP

## Tip podatkov in metode njihovega zbiranja in/ali ustvarjanja

1. Katere podatke boste zbirali oziroma pridobivali in/ali ustvarjali?
2. Na kakšen način boste zbirali oziroma pridobivali in/ali ustvarjali nove podatke in kako boste uporabljali že obstoječe za potrebe vaše doktorske disertacije?
3. Ali boste delali z občutljivimi podatki? Če da, kako boste poskrbeli za etično pridobivanje in/ali ustvarjanje podatkov?

- široka uporabnost jezikovnih podatkov
- pregled že obstoječih zbirk v repozitoriju CLARIN.SI in agregatorju VLO
- pred zbiranjem urediti avtorske pravice, varstvo osebnih podatkov in pogoje uporabe

## Način hranjenja in zaščita podatkov med raziskovalnim delom za doktorsko disertacijo

1. Na kakšen način boste hranili podatke?
2. Če boste delali z občutljivimi podatki, kako boste skrbeli za njihovo varovanje in zaščito?

- strojno berljivi formati zapisa
- sistematična dokumentacija
- varnostno kopiranje
- anonimizacija

## Dolgotrajna dostopnost in hranjenje podatkov

1. Kje oziroma v katerem podatkovnem repozitoriju boste hranili podatke na dolgi rok po zaključku raziskovalnega dela in omogočili njihovo dostopnost?
2. Ali načrtujete za določen čas omejiti dostop do podatkov?

- objava podatkov v repozitoriju CLARIN.SI
- odprta licenca kot privzeta izbira