

# Predstavitev raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN.SI

Tomaž Erjavec

Nacionalni koordinator CLARIN.SI  
Odsek za tehnologije znanja, Institut "Jožef Stefan"

Kako bo mogoče? – posvet o digitalni prihodnosti slovenščine  
2019-09-27

## Kaj je RI?

Naprave, viri in storitve, ki jih uporablja znanstvena skupnost za izvajanje vrhunskih raziskav na svojih področjih.

## Raziskovane infrastrukture ESFRI

- *European Strategy Forum on Research Infrastructures* (ESFRI) je predlagal 15 (2016: 21) RI, nekatere že delujejo kot pravne osebe *European RI Consortium* (ERIC)
- Slovenija sodeluje v 14 ESFRI RI (npr. CERN)
- CLARIN ERIC / CLARIN.SI: Infrastruktura za skupne jezikovne vire in tehnologije (Common Language Resources and Technology Infrastructure)

# CLARIN: Common Language Resources and Technology Infrastructure

- Vizija: digitalni jezikovni viri in orodja za vse (evropske) jezike so dostopni prek enotne prijave za raziskovalce v humanistiki in družboslovju
- Namenjena dolgotrajnemu in obsežnem hranjenju in dostopu do jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske kulturne dediščine
- Nova paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabe in prilagajanja individualnim potrebam



- Sedež na Nizozemskem
- 20 držav članic + 4 opazovalke
- Podporno osebje, odbori za vodenje, delovne skupine
- Večina dela se odvija v okviru nacionalnih konzorcijev



<http://www.clarin.si/>

- Začetek dela v 2014
- Institut " Jožef Stefan"
  - Odsek za tehnologije znanja (E8)
  - Laboratorij za umetno inteligenco (E3)
  - Center za mrežno infrastrukturo (CMI)
- Organiziran kot konzorcij 12 partnerjev:
  - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - 3 raziskovalni inštituti: ZRC SAZU, IJS, INZ
  - 3 društva oz. zavodi: SDJT, Trojina, DDR
  - 2 podjetji: Amebis, Alpineon

Trije stebri:

- Certificiran repozitorij jezikovnih virov in orodij
  - dolgotrajno hranjenje
  - avtentikacija in avtorizacija
  - stalni identifikatorji
  - eksplicitni pogoji uporabe in licence
  - principi FAIR
- Spletne storitve
  - dva konkordančnika (spletne analize korpusov)
  - orodja za označevanje besedil
  - itd.
- Podpora:
  - financiranje priprave virov za vključitev v repozitorij
  - večji projekti: 30.000 EUR letno, 7 projektov v 2018
  - dogodki: JOTA, JT-DH 2018, EURALEX 2018, TSD 2019, ...
  - center znanja za računalniško obdelavo južnoslovanskih jezikov

- Najpomembnejša storitev CLARIN.SI
- Danes 129+ jezikovnih virov, od tega 96 slovenskih: korpusi, slovarji, besedišča, modeli, programi
- Velika večina pod eno od licenc Creative Commons

Repozitorij CLARIN.SI / Prikaz vnosa

Slovenian parliamentary corpus siParl 1.0 (1990-2018)

📄 Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko: **BIBTEX** **CMR**

Pančur, Andrej; Erjavec, Tomaž; Ojsteršek, Mihael; Šorn, Mojca and Blaj Hribar, Neja, 2019, Slovenian parliamentary corpus siParl 1.0 (1990-2018), Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1236>

🔗 Ta vir je integriran tudi v naslednje storitve: **KonText** **noSketch** **Delite:** [f](#) [t](#) [s](#)

**CLARIN.SI Data & Tools**

✍ Avtorji	Pančur, Andrej ; Erjavec, Tomaž ; Ojsteršek, Mihael ; Šorn, Mojca ; Blaj Hribar, Neja
🔑 Identifikator vnosa	<a href="http://hdl.handle.net/11356/1236">http://hdl.handle.net/11356/1236</a>
🌐 URL projekta	<a href="https://github.com/DARIAH-SI/siParl/commit/c6e7942b9fb2199a85e60de6dd30679ce735cf1a">https://github.com/DARIAH-SI/siParl/commit/c6e7942b9fb2199a85e60de6dd30679ce735cf1a</a>
🔗 Demo URL	<a href="http://exist.sistory.si/exist/apps/parla/index.html">http://exist.sistory.si/exist/apps/parla/index.html</a>
📅 Datum objave	2019-05-03
📁 Vrsta	corpus
📊 Velikost	11351 texts, 1083233 utterances, 227896145 tokens
🗣 Jezik(i)	Slovenian

**CLARIN.SI**

🔍 Kaj lahko storite?

**DEPOSIT** **CITE**

📁 Brskanje

> Celoten repozitorij

👤 Moj račun

📄 Prijava

📊 Statistika

📈 Statistika Piwik **BETA**

📄 Splošne informacije

👤 O vnosu v repozitorij

📄 Citiranje

- Pomanjkanje kadra
- Premajhna informiranost skupnosti o obstoju CLARIN.SI (posebej še repozitorija)
- Zakonodaja, ki ovira redistribucijo korpusov



- Namen CLARIN(.SI) je spodbujati raziskave, ki potrebujejo dostop do jezikovnih podatkov
  - digitalna humanistika in družboslovje
  - jezikovne tehnologije
  - vse ostale vede, kjer je jezik pomemben
- Odprt in brezplačen dostop do virov, orodij in storitev
- Slovenski raziskovalci imajo dostop tudi do storitev CLARIN ERIC in drugih nacionalnih konzorcijev CLARIN

# Predstavitev raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN.SI

Tomaž Erjavec

Nacionalni koordinator CLARIN.SI  
Odsek za tehnologije znanja, Institut "Jožef Stefan"

Kako bo mogoče? – posvet o digitalni prihodnosti slovenščine  
2019-09-27