

CLARIN

Tomaž Erjavec

Odsek za tehnologije znanja
Institut " Jožef Stefan"

Odprti raziskovalni podatki v Sloveniji
Maribor, 2019-11-214

Dostopno pod licenco CC BY-SA

Uvod

Kaj so jezikovni viri in tehnologije

- Jezikovni viri:
 - Pisni in govorni korpusi jezika: zbirke besedil
 - izbrane po vnaprej določenih kriterijih
 - opremljene z metapodatki
 - jezikovno označene
 - enovito kodirane
 - Digitalni slovarji
 - Računalniški leksikoni
 - Modeli jezika
- Jezikovne tehnologije:
 - Programi za označevanje besedil
 - Programi za strojno prevajanje
 - Programi za "razumevanje" besedil
 - Govorne tehnologije: sinteza in prepoznavanje jezika

Uporabnost jezikovnih virov

- Humanistika (digitalna humanistika):
 - jezikoslovje: slovaropisje, sociolingvistika, poučevanje jezika, radovednost...
 - digitalni slovarji
 - veliki, avtomatsko označeni in standardno zapisani korpusi jezika
- Računalništvo (jezikovne tehnologije):
 - računalniško procesiranje jezika: učne in testne množice
 - ročno označena besedila (nadzorovano strojno učenje)
 - podporni viri: leksikoni
- Različnost kultur in usmeritev humanistike in naravoslovja
 - slovenščina / angleščina
 - lokalno / globalno
 - mehko / trdo
 - zaprto / odprto

Jezikoslovje in RDA

RDA LDIG: Linguistics Data Interest Group

- Charter: Version 1.0 14th June 2017
- ... disconnect between linguistics publications and their supporting data results in much linguistic research being unreproducible, either in principle or in practice
- Cilji:
 - Common principles and guidelines for data citation and attribution
 - Education & outreach to make linguists more aware of principles of reproducible research and value of data creation methodology, curation, management, sharing, citation and attribution.
 - Ensure greater attribution of linguistic data set preparation within the linguistics profession.

CLARIN.SI

CLARIN ERIC



- ESFRI CLARIN: Common Language Resources and Technology Infrastructure
- CLARIN ERIC (European Research Infrastructure Consortium)
- Sedež na Nizozemskem:
podporno osebje, odbori za vodenje, delovne skupine
- Večina dela se odvija v okviru nacionalnih konzorcijev
- 20 držav članic + 4 opazovalke

CLARIN.SI



- Začetek dela 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
- Dobro sodelovanje z DARIAH-SI/INZ in ADP/FDV

Storitve CLARIN.SI

Trije stebri:

- Certificiran repozitorij jezikovnih virov in orodij
 - dolgotrajno hranjenje
 - avtentikacija in avtorizacija
 - stalni identifikatorji
 - eksplicitni pogoji uporabe in licence
- Spletne storitve
 - dva konkordančnika (spletne analize korpusov)
 - orodja za označevanje besedil
- Podpora:
 - financiranje priprave virov za vključitev v repozitorij
 - večji projekti: 30.000 EUR letno: 7 projektov v 2018, 6 v 2019
 - dogodki: JOTA, JT-DH 2018, EURALEX 2018, TSD 2019, ...
 - CLASSLA K-centre: CLARIN.SI center znanja za računalniško obdelavo južnoslovanskih jezikov

Repozitorij

- Trenutno najpomembnejša storitev CLARIN.SI
- 150+ jezikovnih virov in orodij, od tega prek 100 slovenskih: korpusi, slovarji, besedišča, modeli, programi
- Velika večina pod eno od licenc Creative Commons

Deposit Free and Safe
License of your Choice (Open licenses encouraged)
Easy to Find
Easy to Cite

CLARIN.SI

Search

Advanced Search

Author	Subject	Language (ISO)
Erjavec, Tomaž (51)	TEI (32)	Slovenian (101)
Ljubešič, Nikola (48)	manual annotation (21)	English (22)
Fišer, Darja (18)	lemmatisation (19)	Croatian (18)
Krek, Simon (17)	part-of-speech tagging (19)	Serbian (16)
Dobrovoljc, Kaja (15)	computer-mediated co ... (18)	Bulgarian (7)
... View More	... View More	... View More

What's New

LanguageDescription CLARIN.SI Data & Tools

ELMo embeddings model, Slovenian

Author(s):
UKar, Matej

Description:
ELMo language model (https://github.com/altena/bim-1f) used to produce contextual word embeddings, trained on entire Gizaflida 2.0 corpus (https://vni.cmt.si/oiafida/System/impressumi) for 10 epochs. 1.364.064 most ...

What can you do?

DEPOSIT CITE

Browse

> All of the Repository

My Account

Vnos repozitorija

Repozitorij CLARIN.SI / Prikaz vnosa

ISČI

Slovenian parliamentary corpus siParl 1.0 (1990-2018)

“ Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko:
 BBTEX CMOI

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](#) [g](#)

[CLARIN.SI Data & Tools](#)

✎ Avtorji	Pančur, Andrej ; Erjavec, Tomaž ; Ojsteršek, Mihael ; Šorn, Mojca ; Blaj Hribar, Neja
➔ Identifikator vnosa	http://hdl.handle.net/11356/1236
🌐 URL projekta	https://github.com/DARIAH-SI/siParl/commit/c6e7942b9fb2199a85e60de6dd30679ce735cf1a
🔗 Demo URL	http://exist.sistory.si/exist/apps/parla/index.html
📅 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

Certificiranje

CLARIN center B

- CLARIN ima notranjo certifikacijo repozitoriev
- "B-centre": nacionalni center
- Predpogoj je certificiranje DSA oz. sedaj CTS
- CLARIN.SI repozitorij konec 2015 pridobil DSA
- CLARIN.SI 2016 postal CLARIN B-centre
- Sedaj potrebna recertifikacija
- Trenutno v fazi pridobivanja certifikata CTS
(vloga oddana konec oktobra)
- Po uspešni certifikaciji CTS sledi recertifikacija CLARIN

Izkušnje

- DSA/CTS: naporen proces samoevalvacije, pokriva veliko področij
- CLARIN.SI v ugodni poziciji:
 - CLARIN ERIC že v prvi fazi pripravil dokumente, npr. ToS, CoCo, izhodišča za licence
 - Uporabljamo češko platformo CLARIN/LINDAT, znotraj katere je že poskrbljeno za večino tehničnih zahtev
 - IJS ima dobro računalniško infrastrukturo (nabava opreme, UPS, varnostne kopije)

FAIR

F: Najdljivost

- Bogati in odprti metapodaki: CMDI, CC0, OAI-MHP
- Stalni identifikatorji (handle)
- Intergriran v CLARINov Virtual Language Observatory
- Povezan z OpenAIRE
- Zaveden v katalogih repozitorijev re3data, Open Archives
- Razmeroma dobro indeksiran v Googlu

A: Dostopnost

- Delovanje repozitorija 24/7
- Dostop do virov prek HTTPS
- Avtomatska in redna kontrola delovanja in pravilnosti
- Kjer je potrebna prijava: SAML / AAI (EduGain)

I: Interoperabilnost

- Dobra interoperabilnost metapodatkov (CMDI, DC)
- Podatki shranjeni v različnih formatih
- Tudi orodja imajo različne I/O formate
- V resnici še daleč od prave interoperabilnosti podatkov in orodij

R: Ponovna uporabnost

- Metapodatki razmeroma bogati
- Eksplicitne licence (večina virov pod licencami CC)
- Za vsak vir navedeni avtorji, založnik itd.; povezava med različicami
- Uporaba standardnih formatov za (meta)podatke

Več o CLARIN in FAIR:

de Jong, F.M.G.; Maegaard, Bente; De Smedt, Koenraad; Fišer, Darja; Van Uytvanck, Dieter. (2018) CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 3259 - 3264

Zaključki

Zaključki

- Poslanstvo
 - Odprt in brezplačen dostop do virov, orodij in storitev za (slovenske) raziskovalce in — kjer le mogoče — podjetja
- Izzivi
 - Dostopnost: Avtorske pravice, varovanje zasebnosti
 - Motivacija: Vrednotenje razvitih virov in tehnologij
 - Izobraževanje: Citiranje uporabljenih virov in tehnologij v publikacijah
 - Metodološka in tehnična znanja v raziskovalni skupnosti

CLARIN

Tomaž Erjavec

Odsek za tehnologije znanja
Institut " Jožef Stefan"

Odprti raziskovalni podatki v Sloveniji
Maribor, 2019-11-214

Dostopno pod licenco CC BY-SA