

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut „Jožef Stefan“

UP, 8. 3. 2021

Pregled predavanja

1. Uvod
2. Raziskovalna infrastruktura CLARIN(.SI)
3. Storitve CLARIN.SI
4. Repozitorij CLARIN.SI

Uvod

- Jezikovne tehnologije
 - glavna paradigma: nadzorovano strojno učenje
 - programi so večinoma jezikovno neodvisni
 - potrebujejo pa učne (ročno označene) jezikovne vire
 - + testne podatke
- Empirično podprte jezikoslovne raziskave:
 - temeljijo na realnih (in po možnosti označenih) besedilih
- Označene jezikovne vire potrebujemo za vsak jezik posebej
- Kje lahko dobimo take vire za slovenščino (in druge južnoslovanske jezike)?

Jezikovni viri

1. Korpusi:

- enovito kodirana in dokumentirana zbirka besedil
- besedila (i)zbrana po vnaprej določenih kriterijih
- označena (oblikoskladnja, leme, skladnja, imenske entitete, ...)
- referenčni/specializirani; eno/večjezični; pisni/govorni

2. Leksikoni:

- besedišče jezika
- besede / fraze
- oblikoslovje, skladnja, pomen, povezave, prevodi

3. Modeli:

- podatki za nek program, ki mu omogoči označevanje besedil v nekem jeziku za neko raven označevanja
- npr. CLOG model za lematizacijo slovenščine; Moses model za prevajanje slovenščina → angleščina

Ponovna uporaba

- Klasični pristop:
 - za vsako raziskavo posebej izdelati jezikovne vire
 - viri nedostopni drugim raziskovalcem
- Slabosti:
 - izdelava jezikovnega vira je lahko zelo draga in dolgotrajna, velika izguba časa in denarja, če se to počne večkrat
 - kasnejši raziskovalci ne morejo preveriti ali poboljšati prvih rezultatov
 - vzdržuje se monopol raziskovalcev oz. institucij, ki so vire izdelale
 - viri ne morejo biti uporabljeni pri razvoju produktov

Odprt dostop do rezultatov raziskovalnih projektov

- Brez ovir do publikacij in podatkov:
 - prihranek denarja in časa;
 - izogibanje ponavljanju dela;
 - spodbujanje sodelovanja;
 - večja transparentnost znanstvenega procesa;
 - spodbujanje inovacij
- Zelo močan trend v EU in tudi v Sloveniji
- Problemi pri omogočanju odprtega dostopa do jezikovnih virov:
 - avtorske pravice nad besedili
 - varovanje zasebnosti (tudi pravica do pozabe): GDPR
 - pogoji uporabe spletnih portalov (npr. Twitter)

Raziskovalne infrastrukture

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

[A to Z](#) | [Sitemap](#) | [About this site](#) | [Legal notice](#) | [Cookies](#) | [Contact](#) | [Search](#) | [English \(en\)](#)



RESEARCH & INNOVATION

Infrastructures

[European Commission](#) > [Research & Innovation](#) > [Research infrastructures](#) > [ESFRI](#)





- HOME
- WHAT ARE RIs ?
- MAPS of RIs
- THE EUROPEAN LANDSCAPE
- EU FINANCIAL SUPPORT
- ERIC-LEGAL FRAMEWORK
- SYNERGIES - EU INITIATIVES
- INTERNATIONAL COOPERATION

The ESFRI Roadmap 2016

The [ESFRI Roadmap](#) 2016 identifies the new Research Infrastructures (RI) of pan-European interest corresponding to the long term needs of the European research communities, covering all scientific areas, regardless of possible location.



The 2016 Roadmap consists of 21 ESFRI Projects with a high degree of maturity - including 6 new Projects - and 29 ESFRI Landmarks - RIs that reached the implementation phase by the end of 2015.

The ESFRI Roadmap 2016 was launched on 10 March 2016, in Amsterdam. The event was organized under the [Dutch Presidency](#) by the Royal Netherlands Academy of Arts and Sciences (KNAW) in close cooperation with ESFRI, the European Commission and the Dutch Ministry of Education, Culture and Science. Discussions focussed on strategic roadmapping, long-term sustainability and the socio-economic impact of research infrastructures.

See Event [Agenda](#) and [Live Stream](#)



Highlights



An on-line map to locate the ESFRI infrastructures and their partner facilities
About 400 facilities are part of these distributed



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 uporabnost in prilagajanja individualnim potrebam

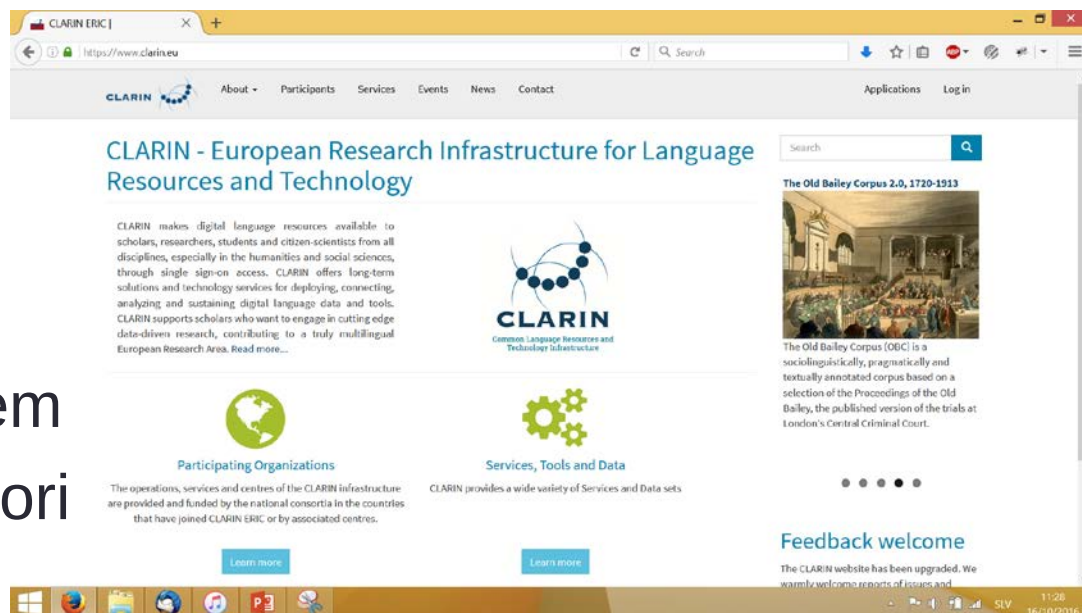


Namen:

- obstoječe vire in orodja dati na voljo v enotni infrastrukturi
- omogočiti svetovalne in učne dejavnosti, kako orodja in vire prilagoditi specifičnim raziskovalnim potrebam
- prispevati k standardizaciji virov in orodij

CLARIN ERIC

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



CLARIN.SI



- <https://www.clarin.si/>
- Začetek dela 2014
- Sedež na IJS:
 - Odsek za tehnologije znanja
 - Laboratorij za umetni inteligenco
 - Center za mrežno infrastrukturo
- Organiziran kot [konzorcij](#) 11 parterjev:
 - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
 - 3 raziskovalni inštituti: ZRC SAZU, IJS, INZ
 - 2 podjetji: Amebis, Alpineon
 - 1 društvo in 1 zavod: SDJT, Trojina

[CLARIN.SI](#)
[REPOZITORIJ](#)
[KONKORDANČNIKI](#)
[STORITVE](#)
[K-CENTER](#)
[DOGODKI](#)
[KONTAKT](#)
[ENGLISH](#)
[CLARIN ERIC](#)



Slovenska raziskovalna infrastruktura za jezikovne vire in tehnologije
Common Language Resources and Technology Infrastructure, Slovenia

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.

Več si lahko preberete na zavihkih na vrhu strani.

OBVESTILA:

Poročila o projektih CLARIN.SI 2020
CLARIN.SI je v letu 2020 ponovil podbudo, začeto v 2018, in objavil projektni razpis za člane svojega konzorcija. Uspešno zaključena projekta (*Razvoj učnega gradiva na korpusu siParl 2.0: Korpusni pristop k raziskovanju parlamentarnega diskurza in izdelava epigrafskega korpusa srednjeveških in zgodnje novoveških napisov na Slovenskem MEMIS*) sta opisana na Projekti CLARIN.SI.

CLARIN.SI

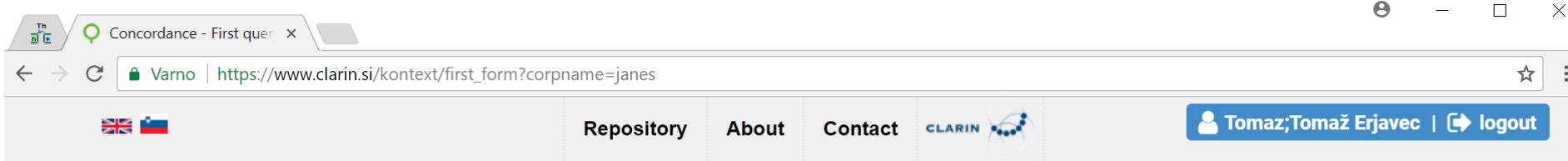


- Repozitorij:
 - Dolgotrajno hranjenje jezikovnih virov in orodij
- Spletne storitve:
 - konkordančniki (analize korpusov)
 - ročno označevanje (izdelava učnih korpusov)
 - delotoki (avtomatizirano označevanje)
- Podpora raziskavam: CLASSLA K-Centre za računalniško obdelavo južnoslovanskih jezikov
- Podpora razvoju jezikovnih virov in tehnologij
 - podpora pripravi virov za vključitev v repozitorij (do 500 EUR)
 - večji projekti: prvič v 2018, 30.000 EUR letno
- Podpora dogodkom:
 - Konference „Jezikovne tehnologije in digitalna humanistika“ (vsake 2 leti, zadnja 2020)
 - 22nd Intl. Conf. on Text Speech and Dialogue, Ljubljana, 11.-13.9., 2019, XVIII EURALEX International Congress, Ljubljana, 2018, ...

Konkordančniki

- KonText + noSketch Engine
- delata lahko z velikimi korpusi (več milijard besed)
- korpusi so lahko bogato označeni:
 - strukture in metapodatki (datum objave, vrsta besedila, spol avtorja, standardnost besedila, ...)
 - atributi besed (oblikoskladenjska oznaka, lema, normalizirana oblika, ...)
- močan poizvedovalni jezik: CQL
- nastavitve pogleda
- raznovrstni izpisi in analize
- CLARIN.SI trenutno ponuja okoli 100 korpusov

KonText (CLARIN-CZ)



[Query](#) [Corpora](#) [Save](#) [Concordance](#) [Filter](#) [Frequency](#) [Collocations](#) [View](#) [Help](#)

Corpus: [Janes \(družbena omrežja\)](#)

Search in the corpus

Corpus:

Query Type: ?

Query: [keyboard](#) | [recent queries](#)

► Specify context

► Specify query according to the meta-information

Konkordance

The screenshot shows the KonText web application interface. At the top, there's a navigation bar with links for Repository, About, and Contact, along with a user profile for Tomaz/Tomaž Erjavec and a logout button. The main header features the 'kon text' logo and a menu with options like Query, Corpora, Save, Concordance, Filter, Frequency, Collocations, View, and Help.

The search results are displayed for the query 'krava' (6,958 hits) in the 'Janes (družbena omrežja)' corpus. The interface shows a list of concordance lines, each with a checkbox, a source link (e.g., 'wiki,comment,slv,negative,T...'), a snippet of text, and the word 'krava' in red. The text snippets are:

- prekmurščini, ker za [per] je internet "indijska sveta
- spet ne smemo. V Novi vasi so navrhu kosti
- o nekem bivšem političnem veljaku, ki je skušal navaditi
- , da bi se prehranjevale z oblanci. In neumne
- lahko preživi (vsaj nekaj časa), tudi tista
- Kot začetnik sem pri opisovanju svoje vasi imel rdečo besedo
- v modro. Pa je začel nastajati članek Domače govedo
- Bil sem ponosen, da sem imel tudi lastno fotografijo
- ponosen, da sem imel tudi lastno fotografijo krav.
- [per], (nekoč so rekli, da če skupaj
- knjigi rekordov. Pa čeprav bi bila prva. thumbSlika
- na govedo, a na sliki po mojem ni cikasta
- majhnimi otroci (po domače rečeno, kot da smo
- ker je [per] vse članke ožigosal kot junca, ali

 The word 'krava' is followed by a comma and a snippet: ', " če tam ne najde, čeprav drugod je

At the bottom, there's a Windows taskbar with various application icons and a system clock showing 16:22 on 07/11/2017.

WebAnno (CLARIN-DE)

The screenshot displays the WebAnno web interface for document curation. The browser address bar shows the URL www.clarin.si/webanno/curation.html?6. The interface has a red header bar with the title "Curation" and a navigation menu including "WebAnno | Home".

Below the header, there are several toolbars:

- Document toolbar:** Includes icons for Open, Re-create, Merge, Prev, Next, Export, and Settings.
- Page toolbar:** Includes navigation buttons (First, Prev, Go to, Next, Last) and a page number input field showing "7".
- Script toolbar:** Includes a button for LTR/RTL.
- Help toolbar:** Includes a button for Guidelines.
- Workflow toolbar:** Includes a button for Finish.

The main content area is divided into three sections:

- Sentences:** A list of sentences on the left, with sentence 7 highlighted in yellow.
- Annotation:** The central area showing the text of 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." and sentence 8: "Anketni vprašalnik (slika 6 – 15) je sestavljen na podlagi najbolj uporabljene metrike za oceno sprejemljivosti tehnologije in informacijskih sistemov (angl. Technology Acceptance Model , TAM) [155] .". Annotations are shown as colored boxes (green for "2TermSlv" and "1TermEng") with arrows indicating relationships (e.g., "(kas. Translation)").
- Actions:** A panel on the right showing the current layer as "kas.BiTerm" and a checkbox for "Forward annotation?". It also indicates "No annotation selected!".

At the bottom, a status bar shows the user "User: andreja_kovacic" and the version "Technische Universität Darmstadt -- Computer Science Department -- WebAnno -- 3.2.2 (2017-07-18 23:57:48, build 61cea1f0f7eda3b57d7c548d0577f166ac2830ce)".

Označevanje besedil

Query

Tagger **Lexicon**

Text

V postopku sta policista ugotovila, da je 45-letni voznik iz Trbovelj vozil s hitrostjo okoli 170 km/h s povsem uničeno pnevmatiko na zadnjem desnem kolesu.

Language Slovenian ▾

Format ☒ Text ☐ TCF

Function ☐ Tag ☐ Lemmatise ☐ Tag + Lemmatise
☐ Tag + Lemmatise + NER
☒ Tag + Lemmatise + Dep Parse

PROCESS **CLEAR**

or

File Browse... No file selected. **REMOVE**

Rezultat

	Surface	Tags	Lemma	Dep parse - gov / func
1.	V	Sl	v	2 / case
2.	postopku	Ncmsl	postopek	3 / nmod
3.	sta	Va-r3d-n	biti	0 / root
4.	policista	Ncmdn	policist	3 / dobj
5.	ugotovila	Vmep-dm	ugotoviti	4 / acl
6.	,	Z	,	5 / punct
7.	da	Cs	da	13 / mark
8.	je	Va-r3s-n	biti	13 / aux
9.	45-letni	Agpmsny	45-leten	10 / amod
10.	voznik	Ncmsn	voznik	13 / nsubj
11.	iz	Sg	iz	12 / case
12.	Trbovelj	Npfpjg	Trbovlje	10 / nmod

Repozitorij



- Zaenkrat najbolj pomembna storitev CLARIN.SI
- Certificiran s strani CLARIN in CTS
- Stalna in varna hramba jezikovnih virov (<https>, Nagios)
- Stalni identifikatorji po sistemu Handle
- Prijava AAI (EduGain)
- Eksplicitni pogoji uporabe (ToS, licenca)
- Etični kodeks (CoCo)
- Žetev metapodatkov (metadata harvesting)
- Večinoma standarden zapis podatkov (XML)

Vstopna stran

Brezplačna in varna hramba

Licenca po vaši izbiri

Preprosto iskanje

Preprosto navajanje vira

CLARIN.SI



Išči

[Napredno iskanje](#)

Avtor

Ljubešić, Nikola (80)

Erjavec, Tomaž (60)

Krek, Simon (31)

Arhar Holdt, Špela (27)

Dobrovoljc, Kaja (26)

... pogledajte več

Ključna beseda

TEI (40)

language model (32)

computer-mediated co ...(30)

manual annotation (29)

part-of-speech tagging (27)

... pogledajte več

Jezik (ISO)

Slovenian (145)

Croatian (33)

English (27)

Serbian (24)

Bulgarian (13)

... pogledajte več

Vstopna stran

Najnovejše

ToolService

CLARIN.SI Data & Tools

The CLASSLA-StanfordNLP model for lemmatisation of standard Slovenian 1.3

Avtor(ji):

Ljubešič, Nikola ; Krsnik, Luka

Opis:

The model for lemmatisation of standard Slovenian was built with the CLASSLA-StanfordNLP tool (<https://github.com/clarinsi/classla-stanfordnlp>) by training on the ssj500k training corpus (<http://hdl.handle.net/11356/1210>) ...

 Ta vnos vsebuje 1 datoteko (37.04 MB).



Publicly Available    

Corpus

CLARIN.SI Data & Tools

Slovenian Twitter hate speech dataset IMSyPP-sl

Avtor(ji):

Kralj Novak, Petra ; Mozetič, Igor ; Ljubešič, Nikola

Opis:

A hand-labeled training (50,000 tweets labeled twice) and evaluation set (10,000 tweets labeled twice) for hate speech on Slovenian Twitter. The data files contain tweet IDs, hate speech type, hate speech target, and ...

 Ta vnos vsebuje 4 datotek(e) (5.19 MB).



Publicly Available    

Brskanje

> Celoten repozitorij

Moj račun

 Prijava

Splošne informacije

 O vnosu v repozitorij

 Citiranje

 Življenjski ciklus vnosa

 Pogosta vprašanja

 O repozitoriju

 Pomoč uporabnikom

RSS Feed

Prijava

Potrebna za vnos novega vira, dostop do zaščitenih virov in za uporabo naprednih funkcij

Repository About Contact CLARIN

Sign in to CLARIN.SI Repository

Login via Your home institution (e.g. university)

- Local authentication
- Clarín.eu website account
- 903 km
- University of Primorska
Slovenia Slovenia
- University of Kentucky
- Brookhaven National Laboratory - SDCC.BNL.GOV
- Iowa State University
- Mengo Hospital
- Okta, Inc
- School center Novo mesto
Slovenia Slovenia
- Northern State University
- Energimuseet
- Slippery Rock University of Pennsylvania

or search for a provider, such as Example University

► Please help, I cannot find my provider

Advanced Search

Subject

- TEI (40)
- language model (32)
- computer-mediated co ... (30)
- manual annotation (29)
- part-of-speech tagging (27)
- ... View More

Uporabniško ime

Uporabniško ime

Geslo

Geslo

Prijava »

Pozabljeno geslo? | Sprememba gesla | Pridobitev E-Identitete

UNIVERSITA DEL TRIESTE
UNIVERZA NA PRIMORSKI

Kako najti zanimive jezikovne vire

- brskanje po jeziku, vrsti vira, ključnih besedah, avtorju itd.
- mehko iskanje
- trenutno na voljo 200+ vnosov

Izdajatelj	
Abu-MaTran project	1
Amebis, d. o. o., Kamnik	1
Centre for Language Resources and Technologies, University of Ljubljana	14
dr. Bruno Breschi Foundation	1
Faculty of Arts, University of Ljubljana	6
Faculty of Computer and Information Science, University of Ljubljana	5
Faculty of Electrical Engineering and Computer Science, University of Maribor	3
Faculty of Electrical Engineering, University of Ljubljana	1
Faculty of Humanities and Social Sciences, University of Zagreb	3
Faculty of Information Studies Novo mesto	4
Faculty of Social Sciences, University of Ljubljana	1
Insight Centre for Data Analytics, National University of Ireland, Galway	1
Institute of Contemporary History	1

Anatomija vnosa

Training corpus ssj500k 2.1



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

BIBTEX

CMDI

Krek, Simon; et al., 2018, *Training corpus ssj500k 2.1*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1181>.



Ta vir je integriran tudi v naslednje storitve:



Delite:



KonText

noSketch

CLARIN.SI Data & Tools

Avtorji Krek, Simon ; Dobrovoljc, Kaja ; Erjavec, Tomaž ; Može, Sara ; Ledinek, Nina ; Holz, Nanika ; Zupan, Katja ; Gantar, Polona ; Kuzman, Taja ; Čibej, Jaka ; Arhar Holdt, Špela ; Kavčič, Teja ; Škrjanec, Iza ; Marko, Dafne ; Jezeršek, Lucija ; Zajc, Anja

Item identifier <http://hdl.handle.net/11356/1181>



URL projekta <http://eng.slovenscina.eu/tehnologije/ucni-korpus>

Datum izdaje 2018-03-16

Zvrst corpus

Velikost 586248 tokens, 27829 sentences, 500293 words

Jezik(i) Slovenian

Anatomija vnosa 2

Ključne besede

[tagging](#)
[dependency treebank](#)
[parsing](#)
[named entities](#)
[tokenisation](#)
[manual annotation](#)
[TEI](#)
[verbal multiword expressions](#)
[semantic role labelling](#)

Zbirke

[CLARIN.SI data & tools](#)

Druge različice

Pokaži vseh različice ▼

[Pokaži vse podatke o v...](#)

Datoteke v tem vnosu



Prenesi navodila za ukazno vrstico



Prevzem vseh datotek v viru (56.28 MB)

Ta vnos je **Publicly Available** z licenco:

Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)



Ime ssj500k-en.TEI.zip
Velikost 28.61 MB
Format application/zip
Opis Corpus encoded in TEI format with annotations in English
MD5 03b7628391fd95b522c1c651bd3998f8

Prevzem datoteke

Predogled



Ime ssj500k-sl.TEI.zip
Velikost 20.59 MB
Format application/zip



Anatomija vnosa 3

dc.contributor.author	Gantar, Polona
dc.contributor.author	Kuzman, Taja
dc.contributor.author	Čibej, Jaka
dc.contributor.author	Arhar Holdt, Špela
dc.contributor.author	Kavčič, Teja
dc.contributor.author	Škrjanec, Iza
dc.contributor.author	Marko, Dafne
dc.contributor.author	Jezeršek, Lucija
dc.contributor.author	Zajc, Anja
dc.date.accessioned	2018-03-16T17:57:20Z
dc.date.available	2018-03-16T17:57:20Z
dc.date.issued	2018-03-16
dc.identifier.uri	http://hdl.handle.net/11356/1181
dc.description	The ssj500k training corpus contains about 500,000 tokens manually annotated on the levels of tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation. About half of the corpus is also manually annotated with syntactic dependencies, named entities, and verbal multiword expressions. About a quarter of the corpus is annotated with semantic role labels. The annotations of the ssj500k corpus follow (1) the MULTEXT-East V5 morphosyntactic specifications for Slovene, http://nl.ijs.si/ME/V5/msd/, (2) the JOS dependency schema, http://nl.ijs.si/jos/bib/jos-skladnja-navodila.pdf, (3) the Janes annotation guidelines for Slovenian named entities, http://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf, and (4) the Guidelines of the

Statistike



Statistics



Statistke Piwik

> View Usage Statistics

Statistike

Showing statistics from 2015-09-14

Skupno št. obiskov

	Ogledi
Emoji Sentiment Ranking 1.0	5654

Skupno število obiskov v določenem letu

	2015	2016	2017	2018	2019
Emoji Sentiment Ranking 1.0	339	1193	1906	2180	36

Top country views

	Ogledi
United States	1659
Slovenija	590
Germany	423
United Kingdom	291
India	224
China	217
Italy	163
France	161

Piwik

✉ Get a monthly report for this item via email

Subscribe

ℹ Click on a data point to summarize by year / month.



Views **1275**



Downloads **18297**

Statistics for the year 2018

Back



Repozitorij: prevzem

- Večina vnosov v CLARIN.SI je CC
- Ostali: AAI prijava, strinjanje s pogoji uporabe
- Citiranje!

Morphological lexicon Sloleks 1.2



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

BIBTEX

CMDI

Dobrovoljc, Kaja; Krek, Simon; Holozan, Peter; Erjavec, Tomaž and Romih, Miro, 2015, *Morphological lexicon Sloleks 1.2*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1039>.



Delite:



CLARIN.SI Data & Tools



Avtorji

Dobrovoljc, Kaja ; Krek, Simon ; Holozan, Peter ; Erjavec, Tomaž ; Romih, Miro



URL projekta

<http://eng.slovenscina.eu/sloleks/opis>



Demo URL

<http://eng.slovenscina.eu/sloleks>



Datum izdaje

2015-06-14

Repozitorij: vnos vira

- Prijava AAI
- Razmeroma enostaven delotok

Item submission



- Zahteve pojasnjene na <https://www.clarin.si/repository/xmlui/page/deposit>
- Izbira licence
- Možnost deljenja vnašanja
- Vnos pregleda eden od urednikov CLARIN.SI
- Možnost embarga podatkov
- Možnost vnosa novejša različice vira

Podatki

- Zahteve pojasnjene na <https://www.clarin.si/repository/xmlui/page/data>
- Idealno: XML po eni od „standardnih“ shem (TEI), za enostavno strukturirane podatke zadošča tabela
- + izvedeni formati + dokumentacija
- ZIP

 Datoteke v tem vnosu



Prenesi navodila za ukazno vrstico



Prevzem vseh datotek v viru (52.35 MB)

Ta vnos je **Publicly Available** z licenco:
Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)



Ime	Sloleks_v1.2.zip
Velikost	26.96 MB
Format	application/zip
Opis	Sloleks in LMF XML format, PoS tags in Slovenian.
MD5	3b15dc1a094e3ff2f4f1ef69702e6f41



 Prevzem datoteke  Predogled

Ime	sloleks-sl.tbl_v1.2.zip
Velikost	12.65 MB
Format	application/zip
Opis	Sloleks in tabular format, PoS tags in Slovenian.
MD5	0c2886e88558df5a6b9cdde4c7e20fb3



 Prevzem datoteke  Predogled

Zakaj deponirati v repozitorij

- Preprečite, da se podatki izgubijo
- Vaši podatki bodo lažje najdljivi
- Če kdo uporabi vaše podatke v objavljeni raziskavi, jih je dolžan citirati
- Spodbujate odprto znanost

Kaj je v repozitoriju?

- Naučeni modeli za analizo jezika (oblikoskladenjsko označevanje, NER, besedne vložitve itd.)
- Ročno označeni korpusi (učne množice)
- Raznovrstni in veliki avtomatsko označeni korpusi (KAS, siParl, IMP itd)
- Leksikalni viri (sloleks, srLex, hrLex,...)
- Slovarji
- Programi



АБУМАТРАН
ABUMATRAN

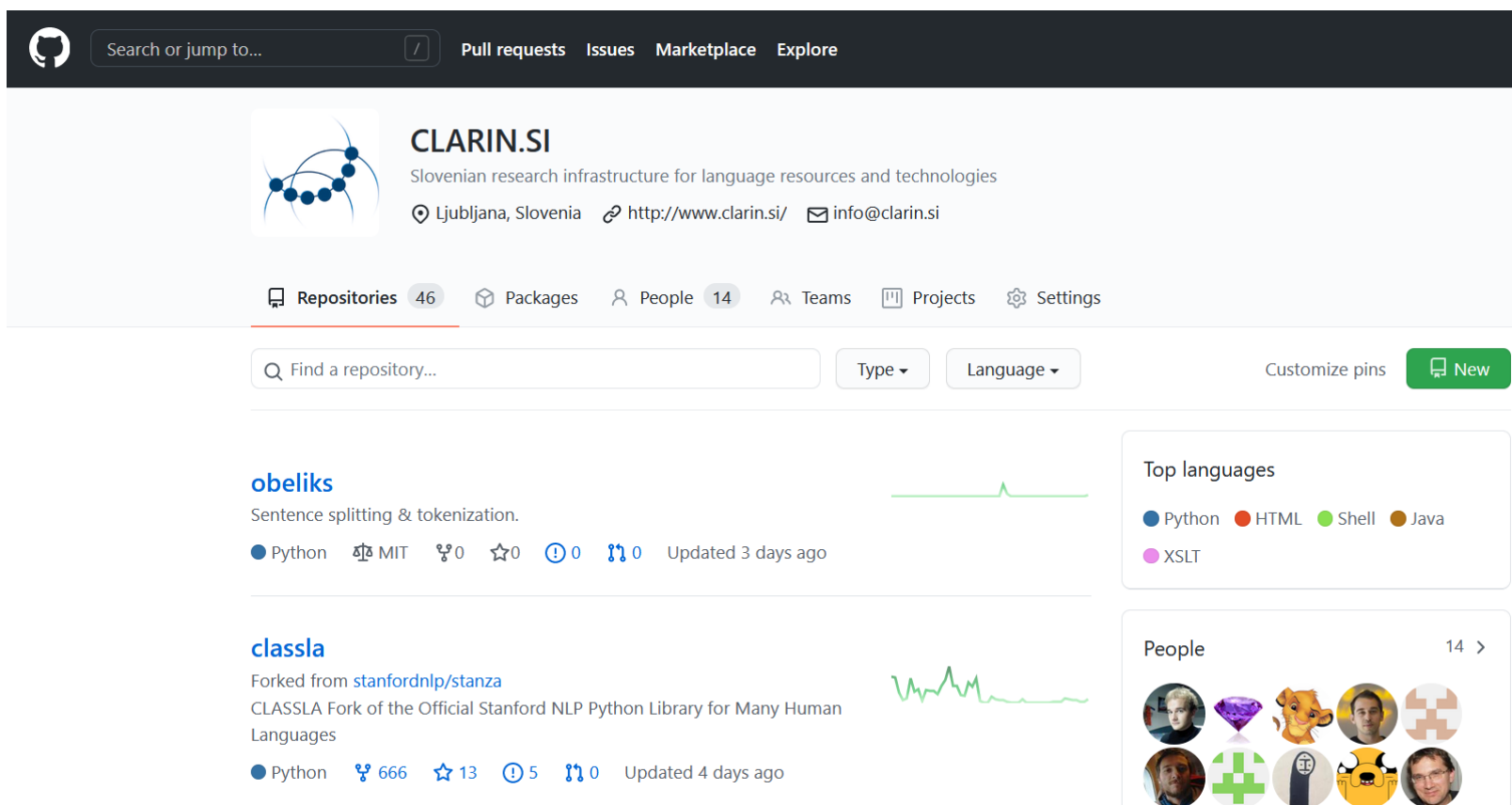


JANES (••)

rsdo

Programska oprema

V repozitoriju sicer so modeli in nekaj programov, vendar se za vzdrževane programske opreme večinoma uporablja GitHub: <https://github.com/clarinsi/>



The screenshot shows the GitHub profile page for CLARIN.SI. The header includes the GitHub logo, a search bar, and navigation links for Pull requests, Issues, Marketplace, and Explore. The profile section features the CLARIN.SI logo, name, and description: "Slovenian research infrastructure for language resources and technologies". It also lists the location (Ljubljana, Slovenia), website (http://www.clarin.si/), and email (info@clarin.si). Below this are tabs for Repositories (46), Packages, People (14), Teams, Projects, and Settings. A search bar for repositories is present, along with filters for Type and Language, and a "New" button. The main content area displays two repositories: "obeliks" (Sentence splitting & tokenization) and "classla" (Forked from stanfordnlp/stanza). The "obeliks" repository shows 0 stars, 0 forks, and 0 issues, updated 3 days ago. The "classla" repository shows 666 stars, 13 forks, 5 issues, and 0 pull requests, updated 4 days ago. On the right, there are sections for "Top languages" (Python, HTML, Shell, Java, XSLT) and "People" (14 contributors).

CLARIN.SI
Slovenian research infrastructure for language resources and technologies
Ljubljana, Slovenia <http://www.clarin.si/> info@clarin.si

Repositories 46 Packages People 14 Teams Projects Settings

Find a repository... Type Language Customize pins New

obeliks
Sentence splitting & tokenization.
Python MIT 0 stars 0 issues 0 pull requests Updated 3 days ago

classla
Forked from [stanfordnlp/stanza](#)
CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages
Python 666 stars 13 forks 5 issues 0 pull requests Updated 4 days ago

Top languages
Python HTML Shell Java XSLT

People 14 >
Contributors: 14 people

Zaključki

- 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 dostop do virov, orodij in storitev