

# Repozitorij CLARIN.SI: uporaba

---

Tomaž Erjavec

Odsek za tehnologije znanja, Institut „Jožef Stefan“

# Vstopna stran

CLARIN.SI repozitorij

← → C Varno | https://www.clarin.si/repository/xmlui/ Prijava

Repozitorij O repozitoriju Kontakt CLARIN

Brezplačna in varna hramba

Licenca po vaši izbiri

Preprosto iskanje

Preprosto navajanje vira

CLARIN.SI

Napredno iskanje

Avtor

- Erjavec, Tomaž (31)
- Ljubešić, Nikola (20)
- Fišer, Darja (13)
- Klubička, Filip (12)
- Krek, Simon (9)
- ... več

Ključne besede

- TEI (27)
- tagging (21)
- lemmatisation (20)
- computer-mediated co ... (16)
- manual annotation (12)
- ... več

Jezik

- Slovenian (62)
- English (15)
- Serbian (9)
- Croatian (6)
- Bulgarian (5)
- ... več

# Vstopna stran

CLARIN.SI repozitorij

Varno | https://www.clarin.si/repository/xmlui/

Najnovejše

**Corpus**

[Tweet code-switching corpus Janes-Preklop 1.0](#)

**Avtor(ji):**  
Reher, Špela ; Erjavec, Tomaž ; Fišer, Darja

**Opis:**  
Janes-Preklop is a corpus of Slovene tweets that is manually annotated for code-switching (the use of words from two or more languages within one sentence or utterance), according to the supplied typology. Words in the ...

[Ta vnos vsebuje 4 datotek\(e\) \(1.28 MB\).](#)

**Publicly Available**

**Corpus**

[Spoken corpus Gos VideoLectures 2.0 \(audio\)](#)

**Avtor(ji):**  
VideoLectures.NET

**Opis:**  
Gos VideoLectures is an add-on to the Gos reference corpus of spoken Slovene (<http://hdl.handle.net/11356/1040>), and covers public academic speech. The Gos VideoLectures corpus contains a selection of public lectures ...

[Ta vnos vsebuje 3 datotek\(e\) \(13.31 GB\).](#)

**Publicly Available**

**CLARIN.SI Data & Tools**

Kaj je na vojo? [Prijava](#)

**DEPOSIT** **CITE**

**Prebrskajte**  
[Celoten repozitorij](#)

**Moj račun**  
[Prijava](#)

**Splošne informacije**

- [O vnosu v repozitorij](#)
- [Navedba vira](#)
- [Življenski ciklus vnosa](#)
- [Pogosta vprašanja](#)
- [O repozitoriju](#)
- [Pomoč uporabnikom](#)

**RSS Feed**  
[RSS 1.0](#)

https://www.clarin.si/repository/xmlui/#

# Prijava

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

The screenshot shows a web browser window with the URL <https://www.clarin.si/repository/xmlui/>. The page has a dark header with tabs for 'Repositorij', 'O repozitoriju', and 'Kontakt'. A red 'Prijava' button is visible in the top right. The main content area features a search bar with a magnifying glass icon and a 'Napredno iskanje' (Advanced search) section. Below these are two columns: 'Avtor' (Author) and 'Ključne besede' (Key words). The 'Avtor' column lists authors like Erjavec, Tomaž (31), Ljubešić, Nikola (20), Fišer, Darja (13), Klubička, Filip (12), and Krek, Simon (9), with an '... več' (more) link. The 'Ključne besede' column lists terms like TEI (27), tagging (21), lemmatisation (20), computer-mediated co..., manual annotation (12), and ... več. A large sign-in overlay titled 'Sign in to LINDAT/CLARIN Repository' is displayed on the right. It asks 'Login via Your home institution (e.g. university)' and lists several options: 'Local authentication' (selected), 'Clarin.eu website account' (with a distance of 929 km), 'Jožef Stefan Institute' (Slovenia, 38 km), 'ŠC Ptuj' (Slovenia, Nearby), 'School center Novo mesto' (Slovenia, Nearby), 'Gymnasium Piran' (Slovenia, Nearby), 'Gological Survey of Slovenia' (Slovenia, Nearby), 'IZUM' (Slovenia, Nearby), and 'Gimnazija Bežigrad high school' (Slovenia, Nearby). There is also a search bar at the bottom of the overlay.

# Kako najti zanimive vire

- brskanje po jeziku, vrsti vira, ključnih besedah, avtorju itd.
- mehko iskanje
- ni težko najti, saj je trenutno samo 80 virov

Faculty of Arts, University of Ljubljana	5
Faculty of Computer and Information Science, University of Ljubljana	4
Faculty of Electrical Engineering and Computer Science, University of Maribor	2
Faculty of Electrical Engineering, University of Ljubljana	1
Faculty of Humanities and Social Sciences, University of Zagreb	2
Faculty of Information Studies Novo mesto	4
Insight Centre for Data Analytics, National University of Ireland, Galway	1
Institute of Contemporary History	1
Jožef Stefan Institute	36
Slovenian Academy of Sciences and Arts	1
Trojina, Institute for Applied Slovene Studies	3
VideoLectures.NET	1
ZRC SAZU	13

# Anatomija vnosa

## Emoji Sentiment Ranking 1.0



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

BIBTEX

CMDI

Kralj Novak, Petra; Smailović, Jasmina; Sluban, Borut and Mozetič, Igor, 2015, *Emoji Sentiment Ranking 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1048>.



Delite:



CLARIN.SI Data & Tools

Avtorji

Kralj Novak, Petra ; Smailović, Jasmina ; Sluban, Borut ; Mozetič, Igor

Demo URL

[http://kt.ijs.si/data/Emoji\\_sentiment\\_ranking/](http://kt.ijs.si/data/Emoji_sentiment_ranking/)

Referenced by

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144296>

Datum izdaje

2015-09-14

Zvrst

lexicalConceptualResource

Velikost

751 entries

Jezik(i)

Albanian , Bulgarian , English , German , Hungarian , Polish , Portuguese , Russian , Serbo-Croatian ,  
Slovak , Slovenian , Spanish , Swedish

Opis

A lexicon of 751 emoji characters with automatically assigned sentiment.

The sentiment is computed from 70,000 tweets, labeled by 83 human annotators  
in 13 European languages.

The process and analysis of emoji sentiment ranking is described in the  
paper: Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. PLoS ONE 10(12):  
e0144296. doi:10.1371/journal.pone.0144296

# Anatomija vnosa 2

EC

Koda projekta: 640772

Ime projekta: DOLFINS

» Ključne besede

sentiment classification emojis unicode

» Zbirke

CLARIN.SI data & tools

Pokaži vse podatke o v...

» Datoteke v tem vnosu

» Prevzem vseh datotek v viru (93.95 KB)

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

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



Ime ESR\_v1.0\_format.txt  
Velikost 783 bajtov  
Format Text file  
Opis Format of the tables.



» Prevzem datoteke

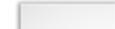
» Predogled

Ime Emoji\_Sentiment\_Data\_v1.0.csv  
Velikost 77.03 KB  
Format Unknown  
Opis CSV table of Emoji Sentiment Ranking



» Prevzem datoteke

Ime Emojitracker\_20150604.csv



# Anatomija vnosa 3

dc.contributor.author	Kralj Novak, Petra
dc.contributor.author	Smailović, Jasmina
dc.contributor.author	Sluban, Borut
dc.contributor.author	Mozetič, Igor
dc.date.accessioned	2015-09-15T17:38:49Z
dc.date.available	2016-04-14T23:00:09Z
dc.date.issued	2015-09-14
dc.identifier.uri	<a href="http://hdl.handle.net/11356/1048">http://hdl.handle.net/11356/1048</a>
dc.description	A lexicon of 751 emoji characters with automatically assigned sentiment. The sentiment is computed from 70,000 tweets, labeled by 83 human annotators in 13 European languages. The process and analysis of emoji sentiment ranking is described in the paper: Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. PLoS ONE 10(12): e0144296. doi:10.1371/journal.pone.0144296
dc.description.provenance	Submitted by Igor Mozetic;Mozetič ( <a href="mailto:igor.mozetic@ijs.si">igor.mozetic@ijs.si</a> ) on 2015-09-15T15:23:10Z No. of bitstreams: 2 Emoji_Sentiment_Ranking_v1.0.csv: 61395 bytes, checksum: fde4f87536275bd7a767339d6fcff486 (MD5) ESR_v1.0_format.txt: 506 bytes, checksum: e1ba7f53f89b74dd3c4e5fcf63a2f09f (MD5)
dc.description.provenance	Approved for entry into archive by Tomaz;Tomaž Erjavec ( <a href="mailto:tomaz.erjavec@ijs.si">tomaz.erjavec@ijs.si</a> ) on 2015-09-15T17:38:49Z No. of bitstreams: 2 Emoji_Sentiment_Ranking_v1.0.csv: 61395 bytes, checksum: fde4f87536275bd7a767339d6fcff486 (MD5) ESR_v1.0_format.txt: 506 bytes, checksum: e1ba7f53f89b74dd3c4e5fcf63a2f09f (MD5)
dc.description.provenance	Made available in DSpace Tomaz;Tomaž Erjavec ( <a href="mailto:tomaz.erjavec@ijs.si">tomaz.erjavec@ijs.si</a> ) on 2015-09-15T17:38:49Z Previous issue date: 2015-09-14 No. of bitstreams: 2 Emoji_Sentiment_Ranking_v1.0.csv: 61395 bytes, checksum: fde4f87536275bd7a767339d6fcff486 (MD5) ESR_v1.0_format.txt: 506 bytes, checksum: e1ba7f53f89b74dd3c4e5fcf63a2f09f (MD5)

# Statistike

Statistics

Statistike Piwik BETA

> View Usage Statistics

## Statistike

Showing statistics from 2015-09-14

### Skupno št. obiskov

	Ogledi
Emoji Sentiment Ranking 1.0	3130

### Skupno število obiskov v določenem letu

	2015	2016	2017
Emoji Sentiment Ranking 1.0	339	1193	1598

### Top country views

	Ogledi
United States	812
Slovenija	444
Germany	217
United Kingdom	181
China	144

# Piwik

CLARIN.SI repozitorij / Emoji Sentiment Ranking 1.0 / Statistike Piwik

BETA

✉ Get a monthly report for this item via email

Subscribe

Period

2017-10-11 - 2017-11-09



👁 Views 260

⬇ Downloads 4417

ℹ Summary



# 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 prednastavljeni 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/ssoleks/opis>

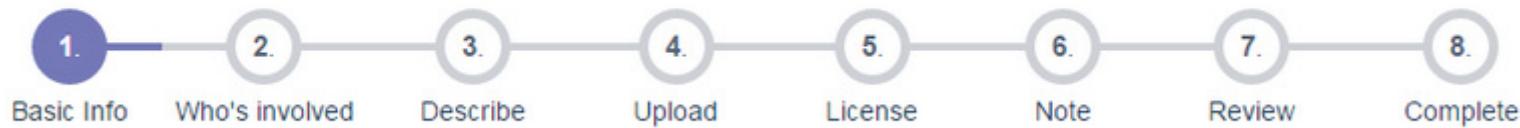
Demo URL <http://eng.slovenscina.eu/ssoleks>

Datum izdaje 2015-06-14

# Repozitorij: vnos

- Prijava AAI
- Razmeroma enostaven delotok

Item submission



- Izbira licence
- Možnost deljenja vnašanja
- Vnos pregleda eden od CLARIN.SI urednikov
- Možnost embarga podatkov
- Možnost vnosa nove različice vira

# Podatki

- XML po eni od „standardnih“ shem
  - Text Encoding Initiative Guidelines (TEI P5)
  - za enostavno strukturirane podatke zadošča tabela
- + izvedeni formati + dokumentacija
- ZIP

 Datoteke v tem vnosu

 [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)



<b>Ime</b>	Sloleks_v1.2.zip	
<b>Velikost</b>	26.96 MB	
<b>Format</b>	application/zip	
<b>Opis</b>	Sloleks in LMF XML format, PoS tags in Slovenian.	

 [Prevzem datoteke](#)  [Predogled](#)

<b>Ime</b>	sloleks-sl.tbl_v1.2.zip	
<b>Velikost</b>	12.65 MB	
<b>Format</b>	application/zip	
<b>Opis</b>	Sloleks in tabular format, PoS tags in Slovenian.	

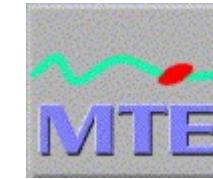
 [Prevzem datoteke](#)  [Predogled](#)

<b>Ime</b>	sloleks-en.tbl_v1.2.zip	
<b>Velikost</b>	12.74 MB	
<b>Format</b>	application/zip	
<b>Opis</b>	Sloleks in tabular format, PoS tags in English.	

 [Prevzem datoteke](#)  [Predogled](#)

# Kaj je zaenkrat v repozitoriju?

- MULTEXT-East
- Viri za učence japonščine
- fi, hr, sr, bs hr-en, fi-en
- Šolar
- sloLeks, hrLex, srLex
- sloWNet
- N-grami
- Emozi
- Slovarji (ZRC SAZU)
- Parlament
- VAYNA



АБУМАТРАН  
ABUMATRAN



JANES



# Učne množice

- sodobna slovenščina: **ssj500k**  
MSDji, leme, imena, skladnja, glagolske fraze (2.0)
- starejša slovenščina: **goo300k**  
posodobljene besede, ~MSDji, leme, imena (faksimile)
- spletni mediji:
  - **Janes-Norm**  
standardizirane besede
  - **Janes-Tag**  
standardizirane besede, MSDji, leme, imena
- govor: **GosVL**  
„fonetična“ in standardizirana transkripcija
- zaledni leksikon: **sloLeks, imp25k**

# Veliki korpusi

metapodatki + avtomatsko označeni: (normalizirane besede), MSDji, leme

- sodobna slovenščina, referenčni korpusi, projekt [SSJ: ccGigaFida](#) (100M), ccKRES (10M)
- zgodovinska slovenščina, projekt [IMP: IMP](#) (18M)
- družbeni mediji, projekt [JANES: Janes-Tweet](#) (160M), [Janes-Forum](#) (50M), [Janes-Blog](#) (35M), [Janes-News](#) (15M), [Janes-Wiki](#) (5M)
- splet (WaC: Web as Corpus)  
[\*\*hrWaC\*\*](#), [\*\*srWaC\*\*](#), [\*\*bsWaC\*\*](#), [\*\*slWaC\*\*](#) (po ~1000M)

# Govorjeni korpusi

- učenje razpoznavanja govora:  
**SNABI** (132 govorcev x 200 stavkov)
- govorjene poizvedbe po letalskih letih:  
**SOFES** (12,536 izjav, 10 ur)
- posneti predavanja:  
**GosVL** (25 predavanj, 10 ur)
- referenčni:  
**Gos** (1000.000 besed, ~~120 ur~~)

# Zaključki

- CLARIN.SI ponuja odprte podatke o slovenskem (in drugih) jezikih, primernih za pedagoške, raziskovalne in tudi razvojne namene
- uporabo podatkov je potrebno citirati
- CLARIN.SI je zelo vesel novih podatkov, če so le:
  - netrivialni
  - razmeroma kvalitetni
  - primerno kodirani
  - dobro opisani
- uporabo vaših podatkov bo potrebno citirati