

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"

Delavnica Odprta znanost
Pravna fakulteta
23. januar 2020



<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 11 partnerjev:
 - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
 - 3 raziskovalni inštituti: ZRC SAZU, IJS, INZ
 - 2 podjetji: Amebis, Alpineon
 - 1 društvo + 1 zavod: SDJT, Trojina

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, 6 v 2019
 - dogodki: JOTA, JT-DH 2018, EURALEX 2018, TSD 2019, ...
 - CLASSLA K-centre: CLARIN.SI center znanja za računalniško obdelavo južnoslovanskih jezikov

- noSketch Engine + KonText
- Obdelava velikih korpusov (več milijard besed)
- Korpusi so lahko bogato označeni:
 - strukture (besedilo, odstavek, termin, ...)
 - metapodatki (datum, vrsta besedila, spol avtorja, standardnost, ...)
 - atributi pojavnic (oblikoskladenjska oznaka, lema, normalizirana oblika, ...)
- Bogat poizvedovalni jezik: CQL
- Raznovrstni izpisi in analize
- RESTful, tj. poizvedbe prek URLjev
- Dostopnih 70 korpusov v 27 jezikih in s prek 15 milijard besed

noSketch Engine

noSketch Engine

Corpus: Gigafida v2.0 DeDup (referenčni, dedupliciran)

Simple query: krava

[Query types](#) [Context](#) [Text types](#)

Query type: simple lemma phrase word character CQL

Lemma: PoS: unspecified

Phrase:

Word form: PoS: unspecified match case

Character:

CQL: Default attribute: word

[Tagset summary](#) [CQL builder](#)

CLARIN.SI
Lexical Computing
2.36.7-open-2.158.8-open-3.105.1

- Odprtokodna različica Sketch Engine
- Prijava ni niti potrebna niti mogoča

The screenshot shows a web browser window with the URL https://www.clarin.si/kontext/first_form?corpname=janes. The page features a navigation bar with links for Repository, About, and Contact, along with a CLARIN logo and a user profile for Tomaz, Tomaž Erjavec with a logout button. The main content area displays the KonText logo and a menu with options: Query, Corpora, Save, Concordance, Filter, Frequency, Collocations, View, and Help. Below the menu, the current corpus is identified as 'Janes (družbena omrežja)'. A search form titled 'Search in the corpus' is shown with the following fields: Corpus (Janes (družbena omrežja)), Query Type (Basic), and Query (krava). There are also expandable sections for 'Specify context' and 'Specify query according to the meta-information', and a Search button.

- Razvil ga je češki CLARIN
- Prijava: shranjene poizvedbe, nastavitve zaslona, podkorpori
- Nima nekaterih funkcionalnosti noSketch Engine

- Zaenkrat najbolj pomembna storitev CLARINa
- Stalna in varna hramba jezikovnih virov (<https>, Nagios)
- Eksplicitni pogoji uporabe (ToS, licenca), etični kodeks (CoCo)
- Standarden zapis metapodatkov: Component Metadata Infrastructure (CMDI) Dublin Core (DC)
- Žetev metapodatkov (metadata harvesting)
- Večinoma standarden zapis podatkov (XML, TEI)
- Trenutno vsebuje 159 virov, od tega 121 slovenskih

Anatomija vnosa, 1

Slovenian parliamentary corpus siParl 1.0 (1990-2018)

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

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: Facebook Twitter Google+

CLARIN.SI Data & Tools

Avtorji Pančur, Andrej ; Erjavec, Tomaž ; Ojsteršek, Mihael ; Šorn, Mojca ; Blaj Hribar, Neja

Item identifier <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

ISGI

CLARIN.SI

Kaj lahko storite?

DEPOSIT CITE

Brskaj

Celoten repozitorij

Moj račun






Prijava

Statistike

Statistika Piwik BETA

- osnovni metapodatki; citiranje; integracija s storitvami; povezave na projekt, demo & objave
- lokalizacija, orodna vrstica, prijava; iskanje, osnovne informacije o repozitoriju, brskanje; Piwik

Anatomija vnosa, 2

 Vrsta	corpus	Splošne informacije Prijava
 Velikost	11351 texts, 1083233 utterances, 227896145 tokens	
 Jezik(i)	Slovenian	
 Opis	<p>The siParl corpus contains minutes of the Assembly of the Republic of Slovenia for 11th legislative period 1990-1992, minutes of the National Assembly of the Republic of Slovenia from the 1st to the 7th legislative period 1992-2018, minutes of the working bodies of the National Assembly of the Republic of Slovenia from the 2nd to the 7th legislative period 1996-2018, and minutes of the the Council of the President of the National Assembly from the 2nd to the 7th legislative period 1996-2018. The corpus comprises over a million speeches or 195 million words. The corpus contains basic meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations.</p> <p>This item comprises three datasets:</p> <ul style="list-style-type: none">- the corpus in TEI (module Transcriptions of speech);- the corpus in TEI with added automatic linguistic annotation: tokenisation, MSD tagging and lemmatisation;- the linguistically annotated corpus in vertical format used by various concordancers, e.g. CWB and Sketch Engine; this format is simpler and smaller but does not contain all the information from the source TEI. <p>A preliminary version of this resource is presented in the paper: Pančur, Andrej, Mojca Šorn and Tomaž Erjavec (2018). "SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession." Darja Fišer and Maria Eskevich and Franciska de Jong (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018. http://rec-conf.org/workshops/lrec2018/W2/summaries/4_W2.html</p>	
 Izdajatelj	Institute of Contemporary History	

- vrsta, velikost, jezik, opis, založnik
- več informacij o repozitoriju


Anatomija vnosa, 3

🔍 Subject(s) parliamentary debates Slovenian Parliament TEI


📁 Collection(s) CLARIN.SI data & tools



Show full item record

📁 Files in this item

 Download instructions for command line

This item is **Publicly Available** and licensed under:
Creative Commons - Attribution 4.0 International (CC BY 4.0)



Name	siParl.zip	
Size	477.97 MB	
Format	application/zip	
Description	Corpus in TEI format	
MD5	08c83cecc1ac42e2cf5a28652aac996d	
Download file	Preview	
Name	siParl-ana.zip	
Size	1.18 GB	
Format	application/zip	
Description	Corpus in TEI format, with linguistic annotations	
MD5	91929e140d2d7fa1426748700d21ebcb	
Download file	Preview	

- ključne besede, vsi metapodatki
- licenca, prevzem podatkov

- **ccGigafida**: Referečni korpus standardne slovenščine vzorčeni odstavki korpusa Gigafida (100 mil. besed)
- **Janes**: Korpusi družbenih omrežij tviti, blogi, komentarji (250 mil. besed)
- **IMP**: Jezikovni viri starejše slovenščine 600 knjig, 1700 - 1918, (17 mil. besed)
- **siParl**: Slovenski parlament zapisniki slovenskega parlamenta 1990-2018 (195 mil. besed).
- **Gos VideoLectures**: Govorni korpus 55 predavanj, 22 ur govora (180,000 besed)

- 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
- Avtentikacija AAI, citiranje (handle)
- Prek 70 korpusov na konkordančnikih
- Prek 150 virov v repozitoriju
- Slovenski raziskovalci imajo dostop tudi do storitev CLARIN ERIC in drugih nacionalnih konzorcijev CLARIN

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"

Delavnica Odprta znanost
Pravna fakulteta
23. januar 2020