Predstavitev CLARIN.SI: repozitorij in konkordančniki

Tomaž Erjavec, IJS

Pregled

1. Uvod
2. Repozitorij CLARIN.SI
3. Konkordačniki
4. Uporaba noSketch Engine

Kaj je CLARIN(.SI)?

- CLARIN: Common Language Resources and Technology Infrastructure
- Evropska raziskovalna infrastruktura
- **Vizija:** digitalni jezikovni viri in orodja za vse (evropske) jezike
- **Cilji:**
  - Dolgotrajno in obsežno hranjenje jezikovnih virov
  - Ohranjanje večjezične evropske kulturne dediščine
  - Sodelovanje pri razvoju in uporabi virov in orodij

CLARIN ERIC

- Sedež na Nizozemskem
- Direktorica dr. Darja Fišer (INZ)
- 26 nacionalnih konzorcijev
  + 2 državi opazovalki + ZDA
- Upravni odbor
- Forum nacionalnih koordinatorjev
- Delovne skupine
  - vključevanje uporabnikov
  - pravna vprašanja
  - standardizacija ipd.
- Večina dela v okviru nacionalnih konzorcijev

Slovenski CLARIN

- Začetek dela 2015
- Sedež na Institutu “Jožef Stefan”
- Organiziran kot konzorcij 12 partnerjev
  a. 4 univerze (LJ, MB, GO, KP)
  b. 4 raziskovalni inštituti (ZRC SAZU, IJS, INZ, ZRS)
  c. 2 podjetji (Amebis, Alpineon)
  d. 1 knjižnica (NUK)
  e. 1 društvo (SDJT)
- Več informacij na https://www.clarin.si
- Vprašanja na info@clarin.si

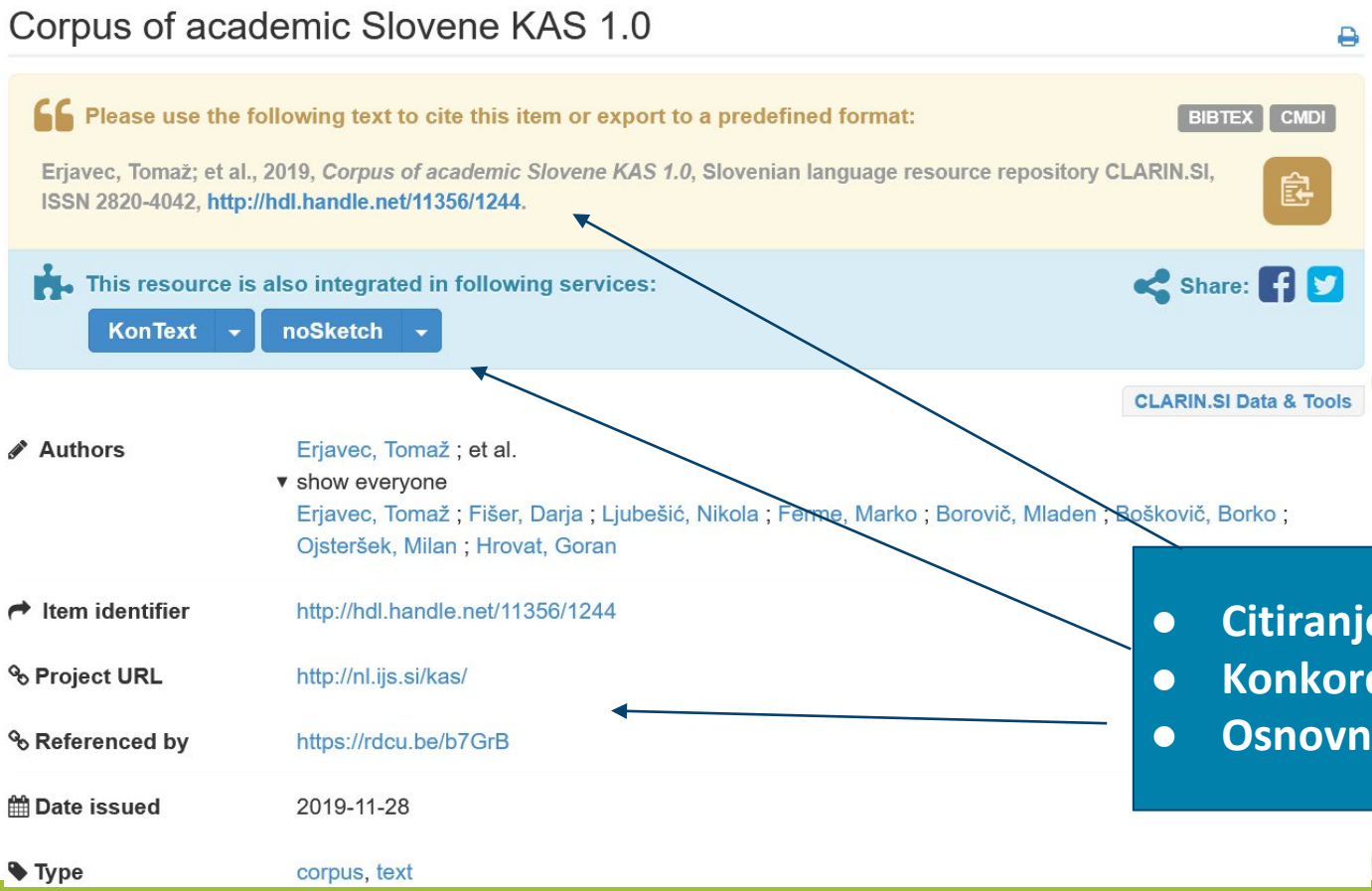
Trije stebri delovanja CLARIN.SI

1. Repozitorij jezikovnih virov in orodij
2. Konkordančniki in druge spletne storitve
3. Podpora raziskovalnim dejavnostim
    - npr. center znanja CLASSLA
    - podpora projektom
    - podpora dogodkom: JT-DH (1998-2026)
    - tečaji, predavanja

Repozitorij CLARIN.SI

- Zaupanja vreden repozitorij
  (Core Trust Seal, CLARIN B-centre)
  - viri in orodja hranjena po načelih FAIR
- Zaenkrat deponiranih cca 700 virov (in orodij) = 9TB
- Podatki o več kot 90 jezikih
- Največ za slovenščino in druge južnoslovanske jezike
- Vrste vnosov:
  - korpusi
  - leksikološki in leksikografski viri
  - jezikovni modeli
  - jezikovna orodja

Repozitorijski vnos 1/3



- Citiranje
- Konkordančnika
- Osnovni metapodatki

Repozitorijski vnos 2/3

Size	82308 texts, 5048551 pages, 1699097710 tokens
Language(s)	Slovenian
Description	The KAS corpus of Slovene academic writing consists of almost 65,000 BSc/BA, 16,000 MSc/MA and 1,600 PhD theses (82 thousand texts, 5 million pages or 1,7 billion tokens) written 2000 - 2018 and gathered from the digital libraries of Slovene higher education institutions via the Slovene Open Science portal (http://openscience.si/). The theses have associated with them significant metadata, while each thesis in the corpus contains its textual body, i.e. without their front and back matter. The body is divided into pages, these into paragraphs, and then into sentences. The sentence tokens are morphosyntactically annotated, words are lemmatised and English-Slovene pairs of term candidates are marked up and linked. The PhD theses in the corpus also have marked-up Slovene monolingual term candidates. The corpus is distributed in the canonical TEI encoding, in the so-called vertical format used by the (no)Sketch Engine and CWB concordancers, and as plain text files. Each format distribution also contains a file with thesis metadata. This repository entry contains the complete corpus; separate entries are available that contain only the PhD theses (KAS-dr: http://hdl.handle.net/11356/1265), the MSc/MA theses (KAS-mag: http://hdl.handle.net/11356/1266) and BSc/BA theses (KAS-dipl: http://hdl.handle.net/11356/1267).
Publisher	Jožef Stefan Institute Faculty of Electrical Engineering and Computer Science, University of Maribor
Acknowledgement	ARRS (Slovenian Research Agency) J6-7094 "Slovene scientific texts: resources and description"

- Velikost
- Jezik(i)
- Opis
- Izdajatelj

Repozitorijski vnos 3/3

Subject(s) PhD theses | MSc/MA theses | BSc/BA theses | academic writing | terminology | TEI

Collection(s) CLARIN.SI data & tools

This item is replaced by a newer submission:
http://hdl.handle.net/11356/1448

List all versions

Show full item record

Files in this item

This item is Academic Use and licensed under:
CLARIN.SI Licence ACA ID-BY-NC-INF-NORED 1.0
Inform Before Use

Name	kas.tei.tar.0.gz
Size	6.31 GB
Format	application/gzip
Description	Corpus in TEI format, slice 0
MD5	83ba9ba74c717c610582c874701c33cb

Download file

- Ključne besede
- Zbirka
- Verzija!
- Licenca
- Datoteke

Objava vira v repozitoriju CLARIN.SI

- Vnos ustvari avtor
- **Urednik pregleda in opozori na potrebne popravke**
- Avtor popravi
- Urednik objavi
- Vir postane viden navzven
  - Google, VLO, DataCite, itd.
  - korpusi običajno integrirani tudi v konkordančnike

Spletne storitve CLARIN.SI

- Konkordančniki
  - NoSketch Engine
  - KonText
- Druga orodja
  - Korpusnik: povzemalnik korpusnih podatkov (CJVT)
  - Senta: Stavčno poenostavljanje in analiza (CJVT)
  - Zbiranje govora: zbiranje posnetkov vsakdanje slovenščine (UM)

Kaj je noSketch Engine?

- Podjetje Lexical Computing je izdelalo konkordančnik **Sketch Engine**
- Sketch Engine ponuja mnogo korpusov v mnogih jezikih
- Uporaba je **plačljiva**

- **noSketch Engine** je odprtokodna različica Sketch Engine
- noSketch Engine ne ponuja nekaterih naprednih funkcij, recimo besednih skic (zato **no**Sketch Engine)
- CLARIN.SI ima svojo instalacijo noSketch Engine s svojimi korpusi

Konkordančniki CJVT in CLARIN.SI

- CJVT: Infrastrukturni center za jezikovne vire in tehnologije Univerze v Ljubljani
- CLARIN.SI: Slovensko vozlišče evropske raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN

- CJVT: izdelal in ponuja iskanje po več ključnih slovenskih jezikovnih virih (slovarji, korpusi)
  - glavni korpusi: Gigafida, Šolar, Gos
- CLARIN.SI: mdr. ponuja iskanje po velikem številu (100+) korpusov s 30+ jeziki
- konkordančniki CJVT ponujajo podobne funkcije kot noSketch Engine, vendar z drugačnim uporabniškim vmesnikom

Jezikovni korpusi

- Korpus:
  velika, urejena, označena in enotno kodirana zbirka besedil
- Uporaba korpusa:
  - prenos, pretvorba in uvoz v želeno orodje (potrebno nekaj znanja programiranja)
  - analiza prek spletnega konkordančnika

Kje se korpusi uporabljajo

- Slovaropisje: nepogrešljiv pripomoček za uvid v uporabo besed
  - Terminologija: luščenje (novih) terminov, kako se uporabljajo, prevodi
- Korpusno jezikoslovje:
  - temelji na podatkih, ne na introspekciji
  - analizira in opisuje dejansko uporabo jezika
  - deskriptivno
  - kvantitativno
  - polno presenečenj :)

Nekaj pomembnejših slovenskih korpusov

Vrste korpusov:

- *Referenčni* proti *specializiranim*
- *Pisni* proti *govornim*
- *Enojezični* proti *večjezičnim*
- *Statični* proti *dinamičnim*

Referenčni korpusi

- Gigafida: najbolj znan slovenski referenčni korpus;
  1 ,1 milijarde besed
- CLASSLA-web.sl 2.0: korpus slovenskega spleta;
  2,2 milijarde besed
- Metafida: korpus narejen iz 39 korpusov, največji korpus slovenskega jezika;
  5 milijard besed
- GOS: referenčni korpus za govorjeno slovenščino;
  2,5 milijonov besed
- SUK: majhen, a ročno označen;
  1 milijon besed

Specializirani korpusi

- Jezik družbenih omrežij: Janes
- Starejša slovenščina: IMP, ELTeC-slv, sPeriodika
- Znanstvene publikacije: KAS, OSS (2 milijardi besed!)
- Parlamentarne razprave:
  siParl, ParlaMint-SI, yu1Parl, Kranjska
- Označene napake/popravki: Šolar, Lektor
- Turistična besedila: TURK
- in mnogo drugih!

Govorni

- Gos: referenčni korpus za govorjeno slovenščino (2,5 milijonov besed)

Večjezični vzporedni korpusi

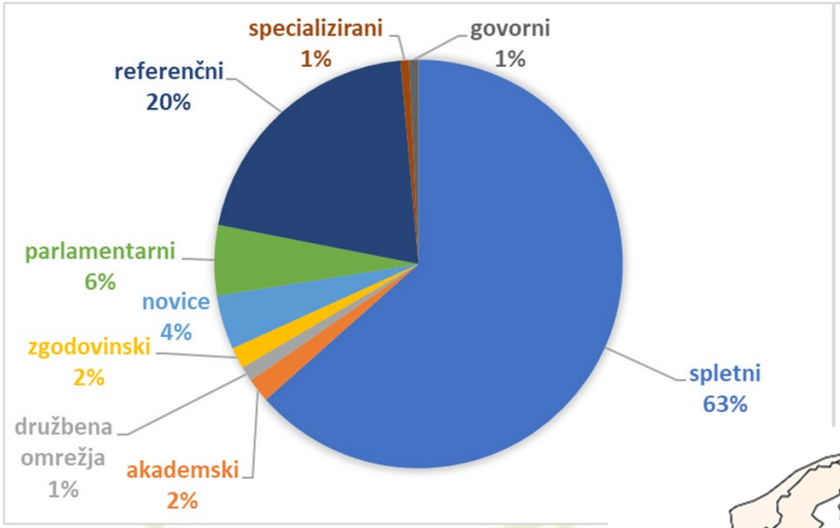
- TRANS5: slovenščina + angleščina
- jaSlo: slovenščina + japonščina
- LeMonde: slovenščina + francoščina
- EU DGT-UD: slovenščina + dosti EU jezikov

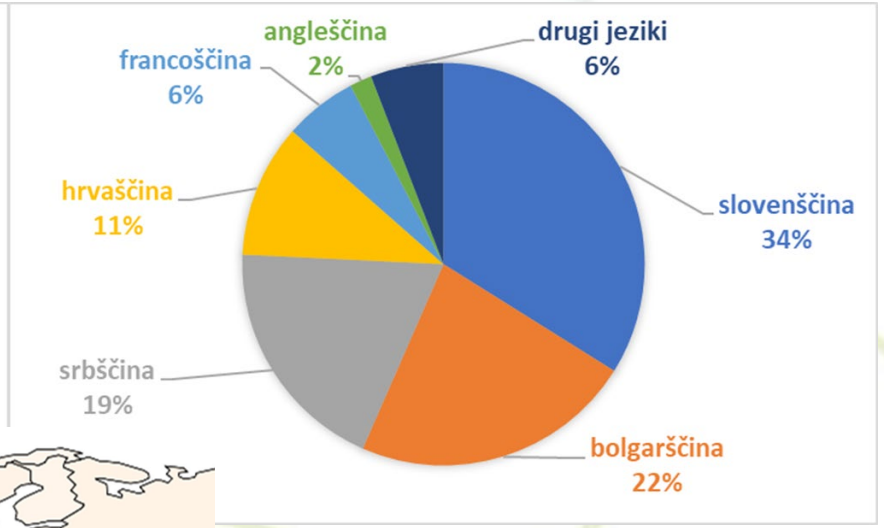
Dinamični (spremljevalni) korpusi

- TRENDI: besedila s spletnih portalov od 2019 dalje
- Trenutno 1 milijarda besed
- Osvežuje se mesečno

Analiza dostopa do korpusov na konordančnikih CLARIN.SI po vrsti, jeziku in državi za 2025

1.12.2024—1.12.2025: 1.011.737 zahtev (45% več kot lani), 2.764 na dan





Vzorec korpusa IMP

```
<text id="FPG_04401-1848" title="Devica Orleanska" author="Schiller, Friedrich"
      date="1848" medium="knjiga" type="umetnostno/gledališka_igra"
      status="prevod">
<p facs="http://nl.ijs.si/ahlib/facs/FPG04401/FPG04401-000.jpg">
<s>
DIVICA          devica          devica          Ncfsn   Sozei
ORLEANSKA       orleanska       orleanski       Agpfsn  Ppnzei
.               .               .               Z       U
</s>
</p>
...
</text>
```

- Strukturne oznake: <text>, <p>, <s>
- Atributi struktur: id, naslov, avtor, …
- Pozicijski atributi: beseda, normalizirana beseda, lema, angl.in slv. oblikoskladenjska oznaka
- **Pomembno**: različni korpusi imajo različne oznake in atribute!

Izdelava korpusa

- Urejanje pravnih vprašanj:
  avtorske pravice, zaščita zasebnosti, pogoji uporabe izvorne platforme
- Zajem in čiščenje besedil
- Zajem in urejanje metapodatkov
- Avtomatsko jezikoslovno označevanje
  - tokenizacija, razdelitev v povedi
  - (normalizacija)
  - oblikoskladenjsko označevanje
  - lematizacija
- Kodiranje korpusa
- Objava korpusa
- Uporaba korpusa:
  - prenos, pretvorba in uvoz v želeno orodje (znanje programiranja)
  - analiza prek konkordančnika

Opozorila

- Izdelati korpus je še vedno kompleksen in dolgotrajen proces
- Zajem besedil je lahko precej slab (PDFji): OSS, sPeriodika
- In tudi sama besedila vsebujejo napake
- Označevanje (npr. lematizacija) se tudi kdaj zmoti:
  - natančnost npr. 95 %: vsaka 20 beseda narobe označena
  - točnost: najdemo samostalnike, ki to niso
  - priklic: ne najdemo samostalnikov, ki niso označeni kot taki
  - napake v označevanju se še posebej vidijo pri leksikalnih poizvedbah

noSketch Engine @ CLARIN.SI

- Brez prijave: https://www.clarin.si/ske/
- S prijavo: https://www.clarin.si/skelog/
  - uporabnik se registrira sam
  - prilagoditev prikaza
  - zgodovina poizvedb
  - **podkorpusi**

Pregled funkcij noSketch Engine

Ključne funkcionalnosti

- pomoč
- opis korpusa
- konkordance:
  metapodatki, razširjeni kontekst, vzorec, premešaj itd.
- prilagoditev prikaza
- frekvenčni seznami
- ključne besede
- kolokacije
- izvoz rezultatov

Korpusnik @CJVT

- https://korpusnik.cjvt.si/
- Hiter in osnovni pregled rabe besed
- Program pošlje poizvedbo na noSkE@CLARIN.SI in povzame rezultate