JTDH 2022

CLARIN.SI kot podpora za raziskovalce

Jakob Lenardič, UL FF
Kristina Pahor de Maiti, UL FF

NAČRT – 1. del

Infrastruktura CLARIN.SI	
Iskanje obstoječih virov	Repozitorij CLARIN.SI
Obdelava in **ustvarjanje** virov	Orodja in repozitorij CLARIN.SI
Ogled in **analiza** virov	Konkordančniki CLARIN.SI in druga orodja
ODMOR	

NAČRT – 2. del

Od iskanja vira do njegove analize **korak po koraku**	Repozitorij CLARIN.SI noSketchEngine siParl ParlaMint-GB
Podpora raziskovalcem	Finance, tehnične zagate in vsebinska vprašanja
VPRAŠANJA	

Infrastruktura CLARIN.SI

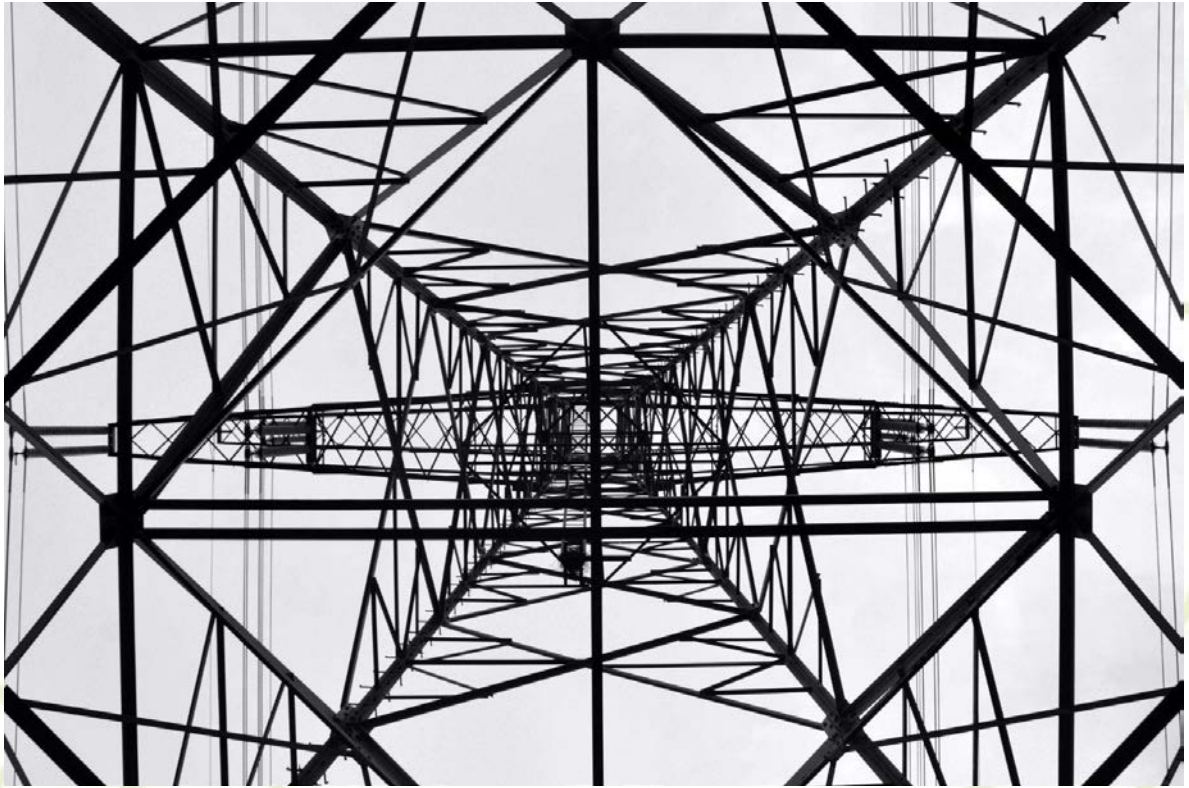


Image by Erik van Dijk

CLARIN: Common Language Resources and Technology Infrastructure

- Raziskovalna infrastruktura
- **Vizija:** digitalni jezikovni viri in orodja za vse (evropske) jezike
- **Cilji:**
  - Dolgotrajno in obsežno hranjenje jezikovnih virov in tehnologij
  - Ohranjanje in podpora večjezične evropske kulturne dediščine
  - Nova paradigma sodelovanja pri razvoju in uporabi virov in orodij

CLARIN ERIC

- Sedež na Nizozemskem
- 22 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

Kaj nudi CLARIN ERIC

- Letna konferenca
  - CLARIN krije stroške za 5 udeležencev na državo in avtorje
- Finančne priložnosti
- Centri znanja
  - K-Centre for Linguistic Diversity and Language Documentation
  - K-Centre for Atypical Communication Expertise
  - K-Centre for South Slavic Languages (CLASSLA)
  - …
- Orodja
  - Virtual Language Observatory (VLO)
  - Language Resource Switchboard (LRS)
- Podporne storitve
  - Resource Families

CLARIN.SI

http://www.clarin.si

- Začetek v 2014
- Nacionalni koordinator: Tomaž Erjavec
- Inštitut “Jožef Stefan”
- Konzorcij 12 partnerjev
  a. 4 univerze
  b. 4 raziskovalni inštituti
  c. 2 društvi
  d. 2 podjetji

Trije stebri CLARIN.SI

1. Repozitorij jezikovnih virov (in orodij)
2. Dva oz. trije konkordančniki in druge spletne storitve
3. Podpora raziskovalnim dejavnostim (vsebinska; finančna; tehnična)

Repozitorij CLARIN.SI

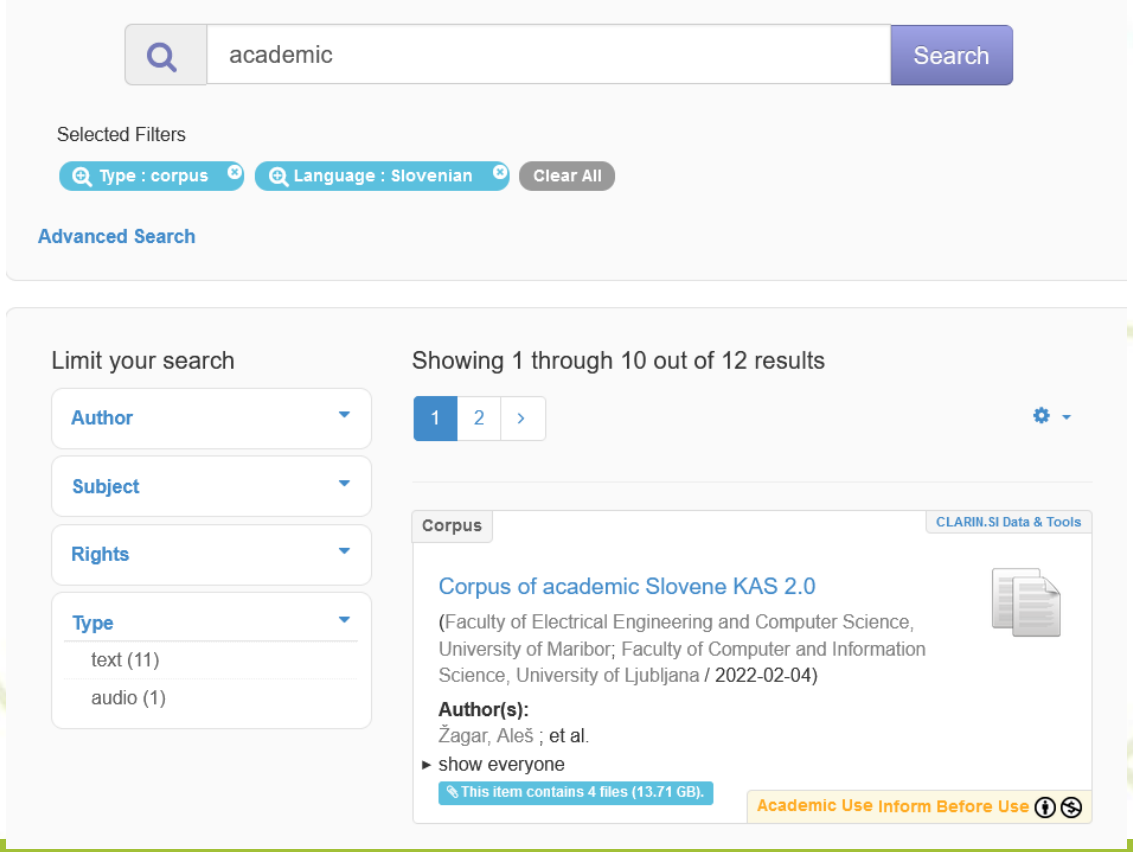
Iskanje obstoječih virov

Image by Ryunosuke Kikuno

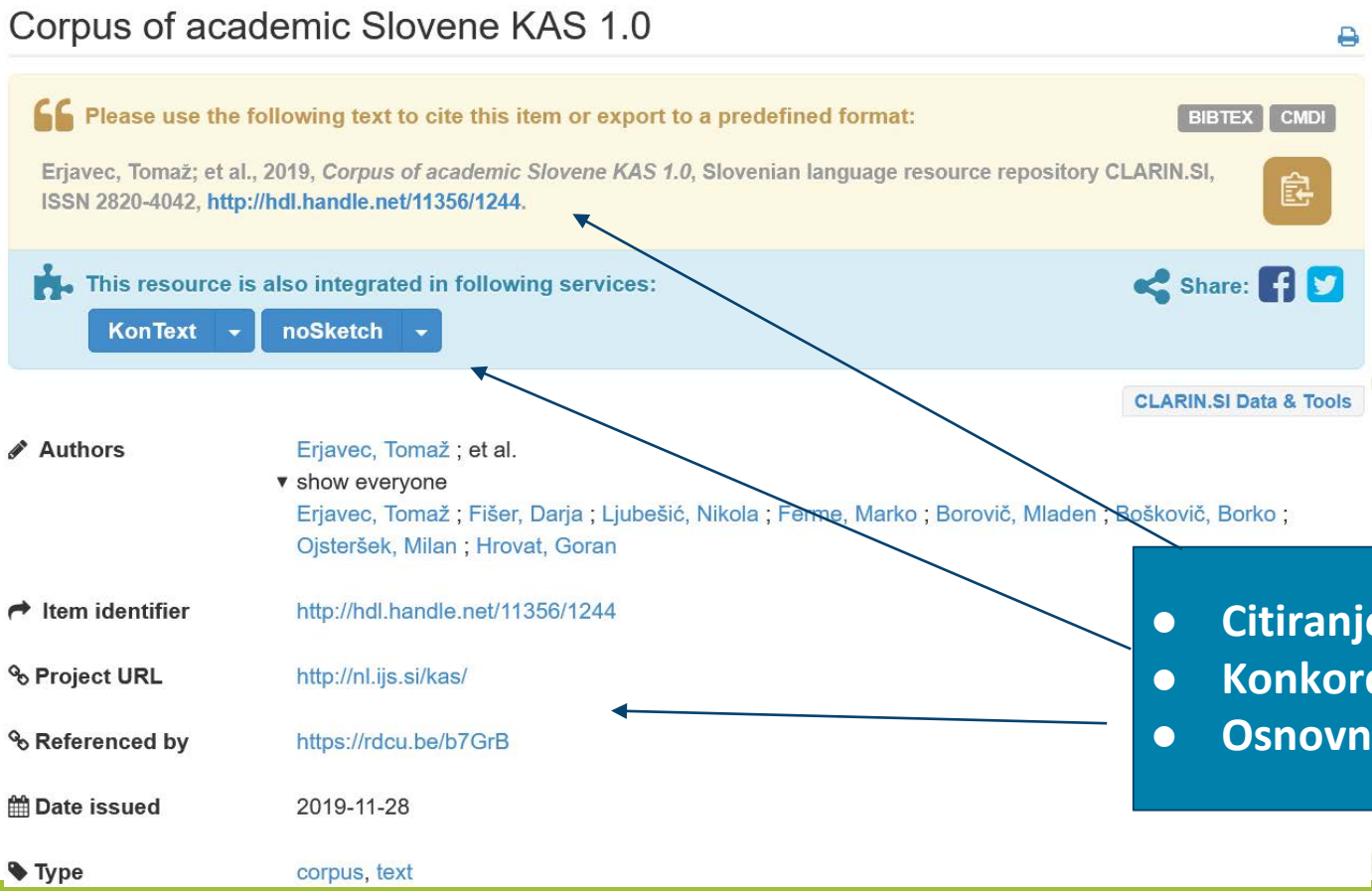
Nekaj dejstev o CLARIN.SI

- Certificiran
  - Core Trust Seal
  - CLARIN B-centre
  - Viri in orodja hranjena po načelih FAIR
- Trenutno deponiranih preko 400 virov (in orodij)
- Podatki o več kot 90 jezikih!
- Vrste vnosov:
  - korpusi
  - leksikološki in leksikografski viri
  - jezikovni modeli
  - jezikovna orodja

Iskanje po repozitoriju



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

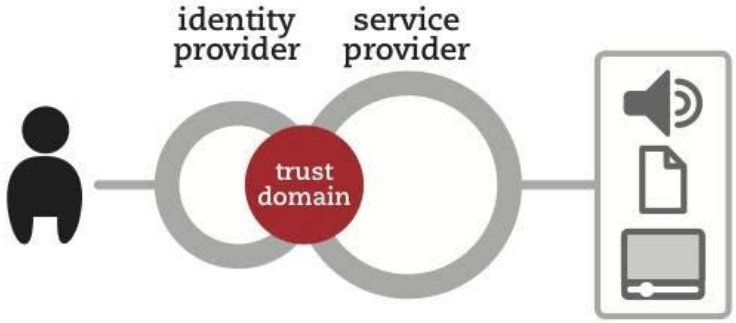
This item is Academic Use and licensed under:
CLARIN.SI Licence ACA ID-BY-NC-INF-NORED 1.0
Inform Before Use

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

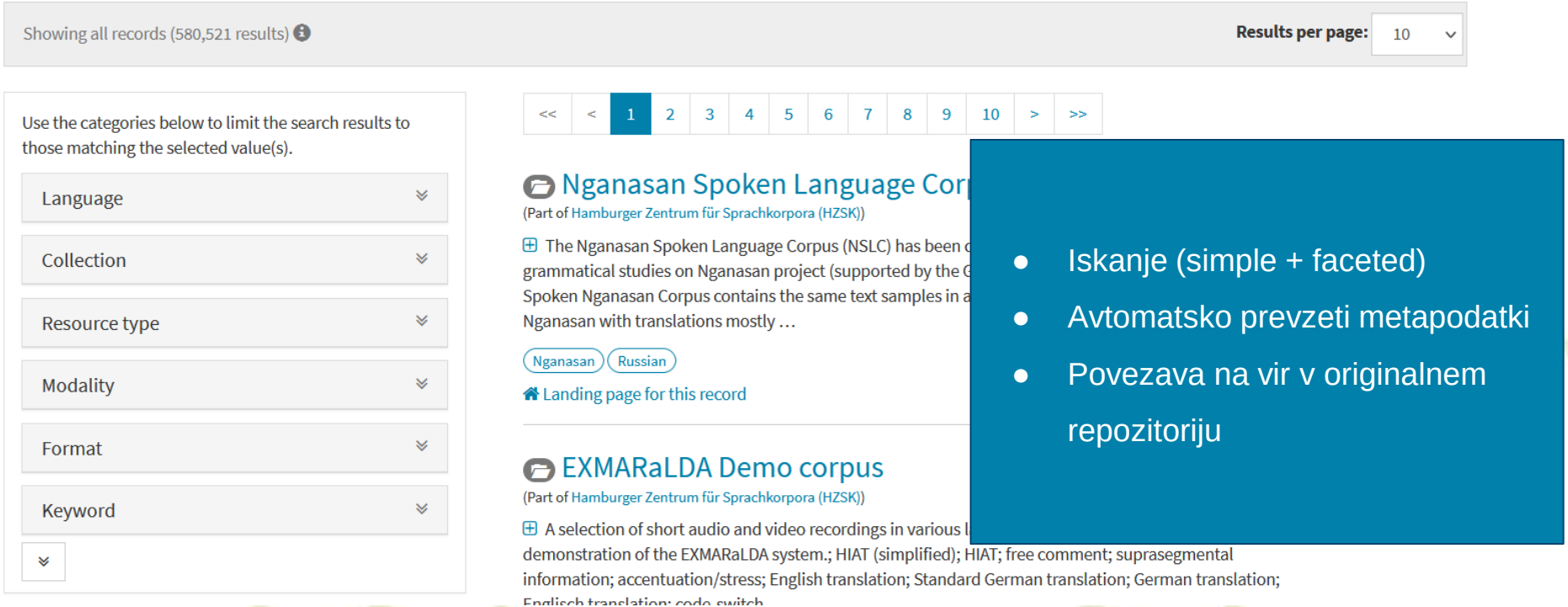
Federated Identity



- “Single sign-on” preko matične institucije
- Lažji dostop do virov na evropski ravni
- Slovenski uporabniki lahko dostopajo do večina CLARIN virov

Virtual Language Observatory (VLO)

https://vlo.clarin.eu



- Iskanje (simple + faceted)
- Avtomatsko prevzeti metapodatki
- Povezava na vir v originalnem repozitoriju

CLARIN Resource Families

https://www.clarin.eu/resource-families

Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Legal corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Obdelava in ustvarjanje virov

Orodja in repozitorij CLARIN.SI

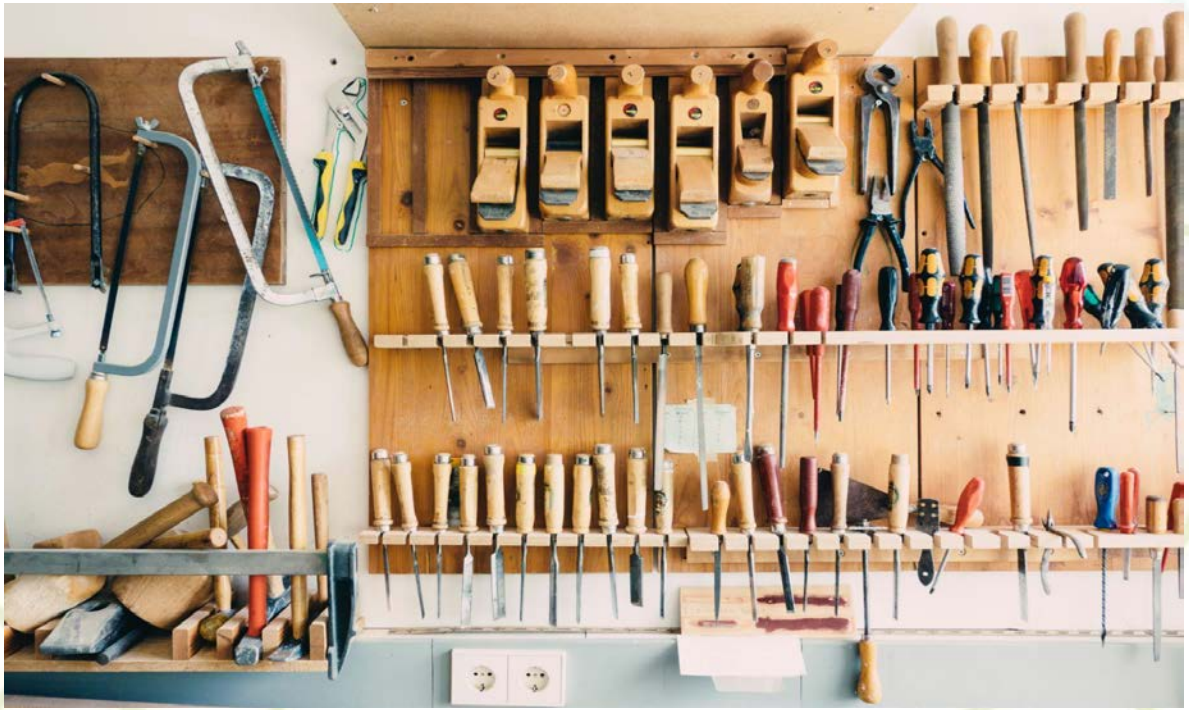


Image by Barn images

Obdelava in označevanje podatkov

- namen
  - obogatitev obstoječih virov
  - izgradnja novih označenih virov

- prednosti označenih besedil
  - **lažje iskanje** (priklic zadetkov po oznakah)
  - **kvalitetnejša in kompleksnejša analiza** podatkov
  - računalniška uporaba (strojno učenje)

- označevanje na različnih ravneh
  - jezikoslovno na ravni besed, stavkov ali celotnih dokumentov
    - **besede/večbesedne enote:** oblikoskladnja, lematizacija, normalizacija, NER
    - **stavki**: odvisnostna skladanja
    - **besedila:** sentiment
  - metapodatkovno

- orodja
  - Webanno – ročno označevanje besedil (podrobneje)
  - ReLDIanno – samodejno označevanje besedil (podrobneje)
  - CLARINSI GitHub organizacija
  - CLARINSI GitLab strežnik

Priprava novega vira

- Zahteve, ki jih mora izpolnjevati kakovosten vir:
  - **VSEBINSKE** zahteve
  - **TEHNIČNE** zahteve
    - načela FAIR: Findable, Accesible, Interoperable, Reusable
      - način zapisa podatkov/format (.xml, .csv., .tsv, .txt, CoNLL-U, .vert, .vrt, TEI xml)
      - podrobneje: predstavitev CTK, predstavitev na konf. Mreža znanja, združenje za raziskovalne podatke RDA
    - načela odprte znanosti – Reliable, Reproducible, Reusable, Relevant
    - Sprotna in sistematična dokumentacija, varnostna kopija: načrt ravnanja z raziskovalnimi podatki
    - priporočila CLARIN.SI glede oblikovanja vira in zahteve za vnos vira v repozitorij CLARIN.SI (možnost tehnične podpore pri CLARIN.SI)
  - **PRAVNE** zahteve

Objava vira v repozitoriju CLARIN.SI

- vnos ustvari avtor
  - clarin.eu
  - repo-help@clarin.si
- prijava preko ponudnika identitete
- urednik pregleda, opozori na popravke in objavi
- vir postane viden navzven
  - Google, VLO, DataCite, arXive itd.
  - korpusi običajno integrirani tudi v konkordančnike

Pravne zahteve

- podrobneje o avtorskih pravicah, licenciranju in varstvu osebnih podatkov – podpora CLARIN
- pomoč pri oblikovanju soglasja za sodelovanje v raziskavi (GDPR)
- vzorec pogodbe CLARIN.SI za prenos avtorskih pravic za namene izdelave in objave korpusa

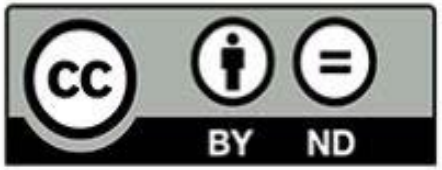
Licence

- opredelitev licence ob vnosu vira
- večina: Creative Commons (CC)
- možno izbrati druge ali ustvariti nove
- v pomoč **OPEN License Selector**

CC BY

CC BY-SA

CC BY-NC-SA



CC BY-ND

CC BY-NC-ND

CC BY-NC

Ogled in analiza virov

Konkordančniki CLARIN.SI in druga orodja

Image by Isaac Smith

Raziskovanje in analiza podatkov

- Konkordančniki
  - NoSketchEngine
  - KonText

- Druga orodja
  - Orange – rudarjenje podatkov
  - Language Resource Switchboard – iskalnik orodij
  - CLARIN.SI repozitorij in VLO

Konkordančniki

- NoSketchEngine
  - prostodostopna različica SketchEngine
  - manjkajo nekatere napredne funkcije
  - stara različica (Bonito)
  - nova različica (Crystal)
  - prijava ni mogoča
- KonText
  - prijava je mogoča
- podrobneje

- Oblikovanje poizvedb
  - prek maske uporabniškega vmesnika
  - poizvedovalni jezik (CQL)
  - regularni izrazi (regex)

- Korpusi
  - cca. 100 korpusov v 33 jezikih (20 mrd besed)
  - raznovrstni in raznojezični
    - največji: metaFida, referenčni (GigaFida), govorni (Gos), spletna besedila (Janes, Tweet-sl), vzporedna (EU-DGT, TRANS5, LeMonde), specializirana (konji, starejša slovenščina 18.–20.st (IMP), filmska kritika, šolski spisi, ELIZA, parlamentarni (ParlaMint, siParl), itd.), drugi jeziki (ang, jap, hr, srb, mak, črnogor)
  - podrobneje

Prikaz uporabe repozitorija CLARIN.SI in konkordančnika noSketchEngine na primeru korpusov siParl in ParlaMint-GB

Image by Isaac Smith

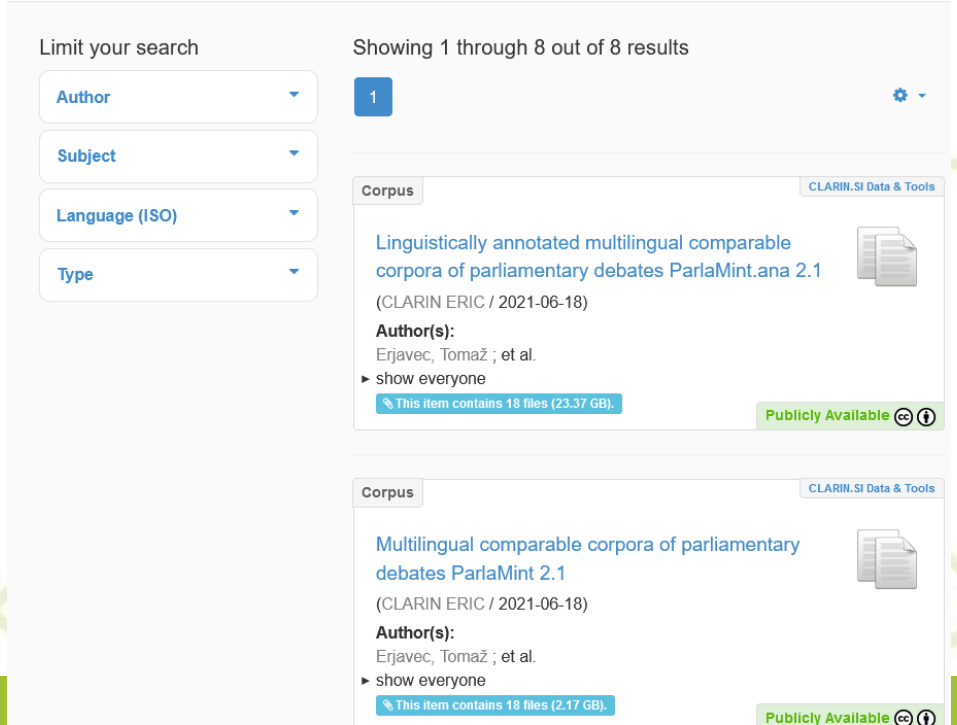
Iskanje v repozitoriju

Nekaj iskalnih nizov:

1. “parallel corpus”
2. “parallel AND corpus”
3. “subject:parallel AND type:corpus”
4. “subject:parallel AND “subject:word sense AND type:corpus”

Parlamentarni korpusi

“type:corpus AND parliament*”



noSketchEngine

- ključne funkcionalnosti
  - opis korpusa
  - konkordance in prilagoditev prikaza
  - izdelava podkorpusa
  - frekvenčni seznami
  - ključne besede
  - kolokacije
- podrobneje → učno gradivo

Značilnosti korpusa

- Splošne informacije (dokumentacija, nabor oznak)
- Zadetki (število pojavnic/tokens vs. besed)
- Leksikon (jezikovne oznake)
- Strukture in atributi (strukture oznake in metapodatki)
- Podkorpusi

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 Council of the President of the National Assembly from the 2nd to the 7th legislative period 1996-2018. The corpus comprises over 10 thousand sessions, one million speeches or 200 million words. The corpus contains meta-data about the speakers, a typology of sessions etc. and structural, editorial and linguistic annotations. The corpus is encoded according to the Parla-CLARIN schema (https://github.com/clarin-eric/parla-clarin). Each mandate is in one directory, and each session in one file.

This item comprises the following datasets:
1. source DARAH-SI Parla-CLARIN encoded corpus;
2. linguistically annotatated Parla-CLARIN encoded corpus: tokenisation, MSD tagging, lemmatisation, Universal Dependencies features and syntactic parses, named entities;

http://hdl.handle.net/11356/1300

Enostavno iskanje

- konkordance hiša
- možnosti prikaza

Izdelava podkorpusa

- Prek vmesnika
  - Išči → Izdelaj novega (klik-klik, nujna preverba)
- Prek konkordanc
  - [lemma="ženska"] within <speech gender="F" & role_en="MP" & type_en="Regular speaker"/>

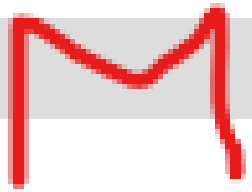
Frekvenčni seznam

- najproduktivnejša politična stranka
  - št. pojavnic
  - št. govorov
- najpogostejši samostalniki pri poslankah
  - concordances → frequency → lemma (lc) → results

Ključne besede

- Mandat1-Ženske vs. Mandat1-Moški
- lemma
- [a-z].*
- ogled konkordanc

porodnišnica	rudnik
deti	proporcionalen
zavoljo	premog
konjerejskovisturističen	kaseta
umrljivost	tule
porod	kolesarski
prezgodnji	termoelektrarna
učinkovina	dama
porodnica	steza
učenka	električen
medicinec	gasilski



Kolokacije

- lema ženska
  - VsiMandati-Ženske vs. VsiMandati-Moški
  - frekvenca leme!
- kolokacije → lemma (lc)
  - korpus M vs. korpus F
- prikaz konkordanc

P | N	moški
P | N	nasilje
P | N	ženska
P | N	nad
P | N	zastopanost
P | N	samski
P | N	enak
P | N	enakost
P | N	delež
P | N	participacija
P | N	politika

Ž

P | N	moški
P | N	samski
P | N	ženska
P | N	nasilje
P | N	zastopanost
P | N	otrok
P | N	enakopravnost
P | N	enakost
P | N	spol
P | N	oploditev
P | N	ploden

M

Karakterizacija žensk in moških – siParl

- [lemma="ženska|dekle|mati" & ud_dep="nsubj"] [ud_dep="cop"]
  - konkordance
  - frekvenčni seznam → lemma (lc) & 1R/1D
- [lemma="moški|fant|oče" & ud_dep="nsubj"] [ud_dep="cop"]
  - konkordance
  - frekvenčni seznam

Karakterizacija žensk in moških – ParlaMint-GB

- [lemma="woman|girl|mother|female .*" & dep="nsubj"] [dep="cop"]
  - konkordance
  - kolokacije → lemma (lc) + razpon: 0-3
- [lemma="man|boy|father|male .*" & dep="nsubj"] [dep="cop"]
  - konkordance
  - kolokacije

parlamint21_gb	parlamint21_gb
1297 hits for WOMAN etc.	623 hits for MAN etc.
underrepresented	breadwinner
pregnant	island
likely	labourer
twice	vicar
frightened	imam
unaware	likely
reluctant	merchant
polish	twice
concentrated	physically
unhappy	dead
desperate	mother
unable	continent
dead	entire
capable	soldier
fearful	innocent
less	perpetrator
absent	graduate
active	victim
angry	stronger
disadvantaged	father

	WOMAN (%)	MAN (%)
positive	25.93	20.83
negative	**66.67**	25.00
profession	7.41	**50.00**

Podpora raziskovalcem

Image by Ralph Kayden

CLASSLA

- Center znanja za južnoslovanske jezike
- Pogosta vprašanja
- Helpdesk: helpdesk.classla@clarin.si
- Spletne storitve (*CLASSLA označevalni cevovod*)
- Delavnice

Razpis za projekte CLARIN.SI

- letni razpis CLARIN.SI, namenjen izdelavi ali nadgradnji virov ali storitev, ki pripomorejo k uresničevanju infrastrukture CLARIN.SI
  - izdelava ali nadgradnja virov, spletnih storitev ali programske opreme
  - organizacija izobraževalnih dogodkov oz. priprava izobraževalnih gradiv
  - raziskave, ki uporablja obstoječe vire ali storitve CLARIN.SI
- rok za oddajo prijave običajno spomladi
- v 2022 sredstva v višini 2.000–10.000 EUR

CLARIN ERIC finančna podpora

- Ogromno priložnosti:

  https://www.clarin.eu/funding

- Mobility Grants!

Zaključek

Image by Markus Spiske

Vprašanja

info@clarin.si

Image by Camylla Battani

Hvala za pozornost.

Image by Eric Brehm