The CLARIN.SI repository

Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute

FRI, 29th April, 2020

Repository

- Currently the most important service of CLARIN.SI
- Long-term and safe archiving of language resources (https, Nagios)
- Explicit terms of use (terms of service, licence)
- Ethical codex (Code of conduct)
- Most resources in standard encoding (Unicode, XML)
- Certified as CLARIN Centre B
  - Digital Seal of Approval, DSA
  - Currently being (re)certified for Core Trust Seal, CTS

Repository platform

- Based on the DSpace platform, developed for open digital repositories
- DSpace modified for use by CLARIN repositories: CLARIN/DSpace developed by Czech CLARIN
- Used by Czech, Italian, Norwegian, Polish, and Slovenian CLARIN
- Maintenance on GitHub, Slovenian fork on GitLab

Metadata

- Standard encoding of metadata:
  - Component Metadata Infrastructure (CMDI)
  - Dublin Core (DC)
- Metadata always CC0
- Meta-data harvesting:
  - CLARIN Virtual Language Observatory (VLO)
  - also:

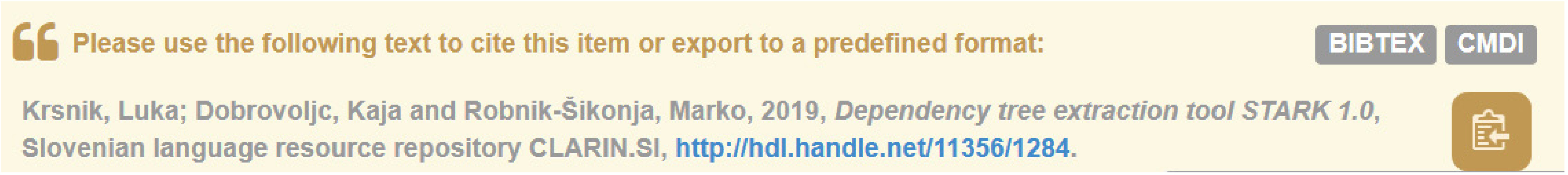
OPEN ARCHIVES

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
http://doi.org/10.17616/R33M0X
CLARIN.SI repository

OpenAIRE

Permanent identifiers

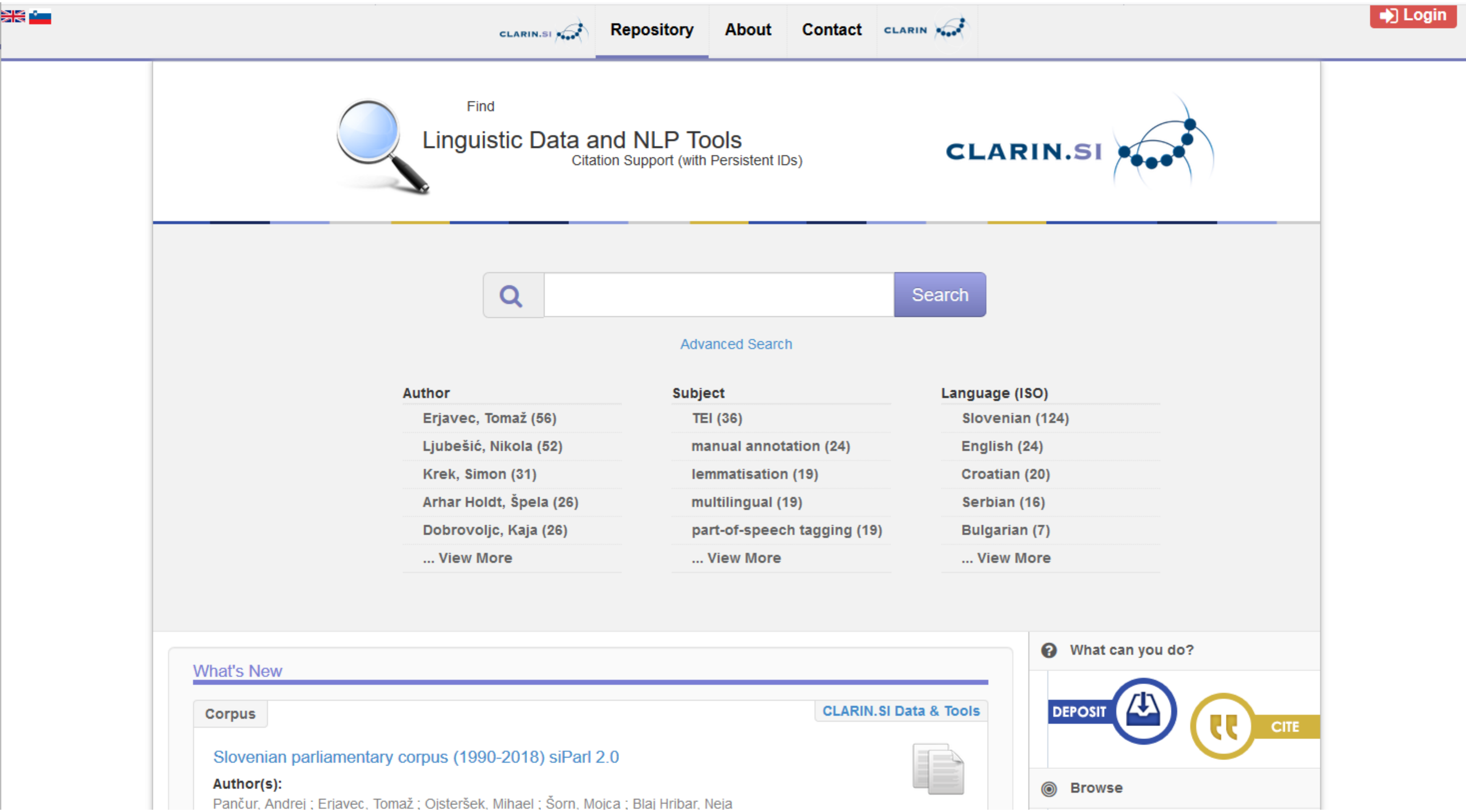
- Each repository entry is assigned a PID
- So, even if the repository platform is changed, the PIDs stay the same (but need to be reconfigured)
- Best known solution: DOI
- CLARIN used the Handle system
- **http://hdl.handle.net/11356/1044** → https://www.clarin.si/repository/xmlui/handle/11356/1044
- Important for citation of repository items:



Please use the following text to cite this item or export to a predefined format:

Krsnik, Luka; Dobrovoljc, Kaja and Robnik-Šikonja, Marko, 2019, *Dependency tree extraction tool STARK 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1284.

Top level page



Find

Linguistic Data and NLP Tools

Citation Support (with Persistent IDs)

Advanced Search

Author	Subject	Language (ISO)
Erjavec, Tomaž (56)	TEI (36)	Slovenian (124)
Ljubešić, Nikola (52)	manual annotation (24)	English (24)
Krek, Simon (31)	lemmatisation (19)	Croatian (20)
Arhar Holdt, Špela (26)	multilingual (19)	Serbian (16)
Dobrovoljc, Kaja (26)	part-of-speech tagging (19)	Bulgarian (7)
... View More	... View More	... View More

What's New

Corpus

CLARIN.SI Data & Tools

Slovenian parliamentary corpus (1990-2018) siParl 2.0

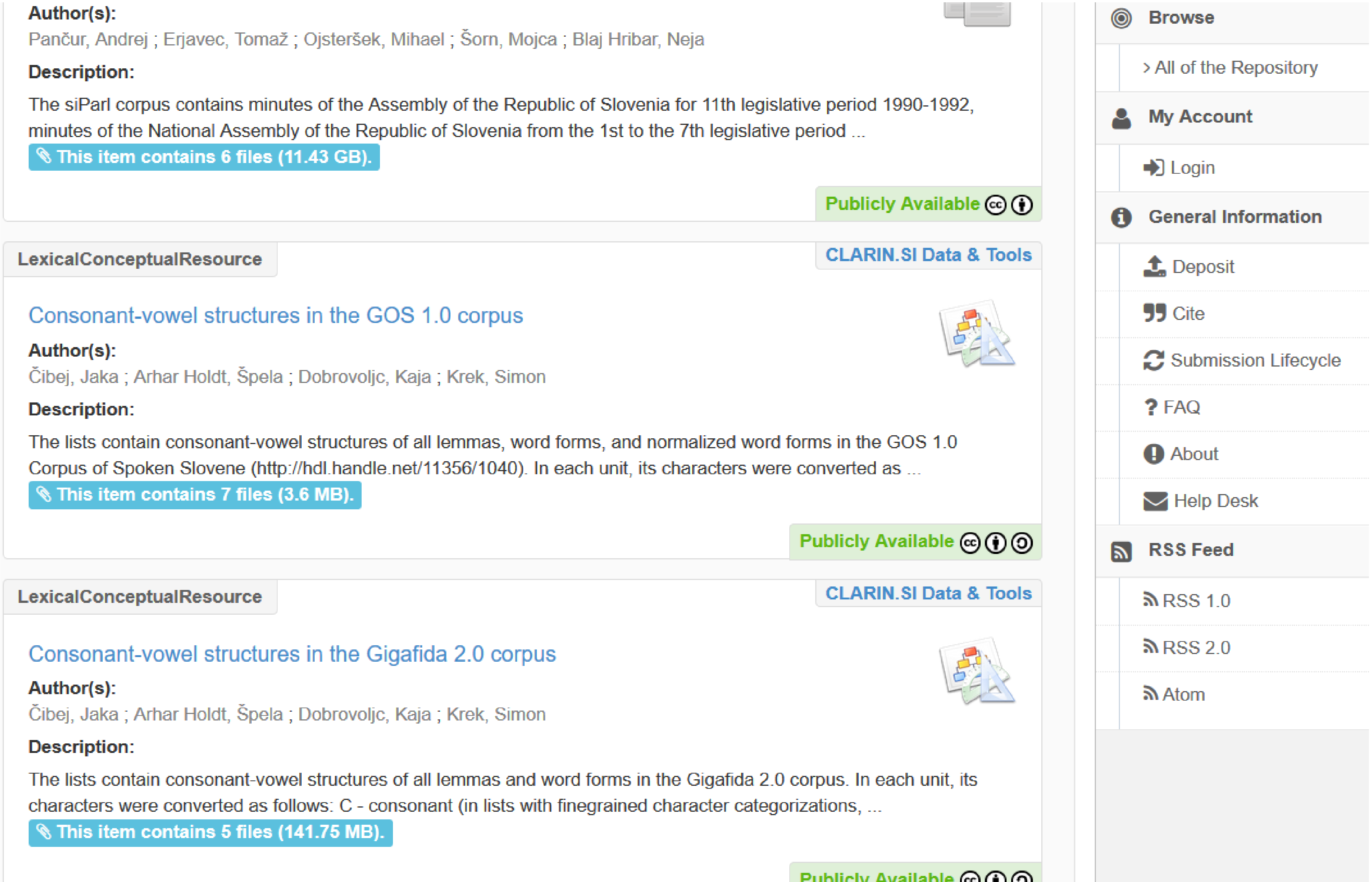
Author(s):

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

What can you do?

Browse

Top level page



Author(s):
Pančur, Andrej ; Erjavec, Tomaž ; Ojsteršek, Mihael ; Šorn, Mojca ; Blaj Hribar, Neja

Description:

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 ...

This item contains 6 files (11.43 GB).

Publicly Available

LexicalConceptualResource

CLARIN.SI Data & Tools

Consonant-vowel structures in the GOS 1.0 corpus

Author(s):
Čibej, Jaka ; Arhar Holdt, Špela ; Dobrovoljc, Kaja ; Krek, Simon

Description:

The lists contain consonant-vowel structures of all lemmas, word forms, and normalized word forms in the GOS 1.0 Corpus of Spoken Slovene (http://hdl.handle.net/11356/1040). In each unit, its characters were converted as ...

This item contains 7 files (3.6 MB).

Publicly Available

LexicalConceptualResource

CLARIN.SI Data & Tools

Consonant-vowel structures in the Gigafida 2.0 corpus

Author(s):
Čibej, Jaka ; Arhar Holdt, Špela ; Dobrovoljc, Kaja ; Krek, Simon

Description:

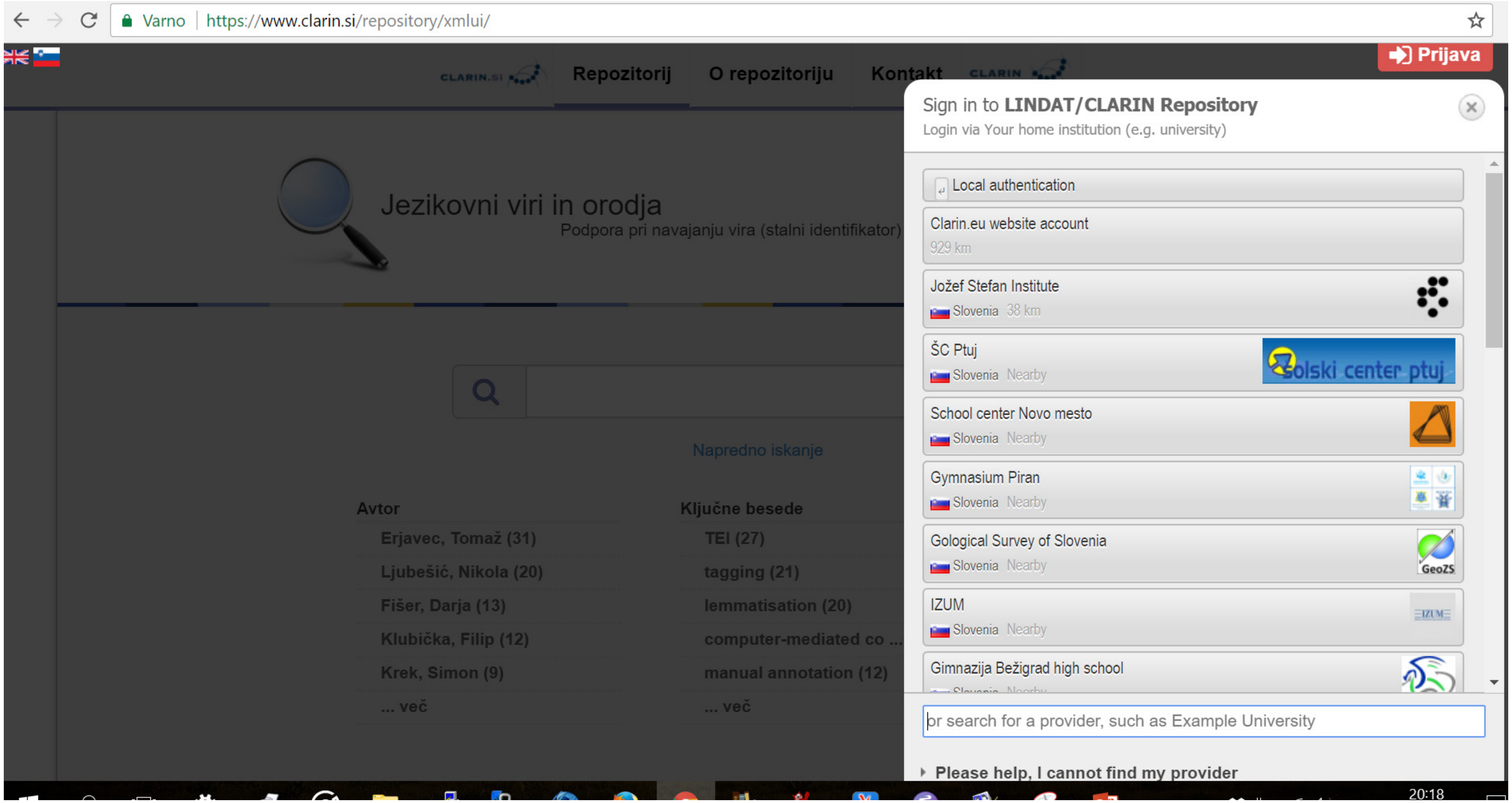
The lists contain consonant-vowel structures of all lemmas and word forms in the Gigafida 2.0 corpus. In each unit, its characters were converted as follows: C - consonant (in lists with finegrained character categorizations, ...

This item contains 5 files (141.75 MB).

Publicly Available

Log-in

- Necessary for making a new repository item, for accessing non-CC items and for using some advanced functions
- For those without EduGain, CLARIN ERIC also gives accounts



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

Jezikovni viri in orodja

Podpora pri navajanju vira (stalni identifikator)

Napredno iskanje

Avtor	Ključne besede
Erjavec, Tomaž (31)	TEI (27)
Ljubešić, Nikola (20)	tagging (21)
Fišer, Darja (13)	lemmatisation (20)
Klubička, Filip (12)	computer-mediated co ...
Krek, Simon (9)	manual annotation (12)
... več	... več

Sign in to **LINDAT/CLARIN Repository**

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

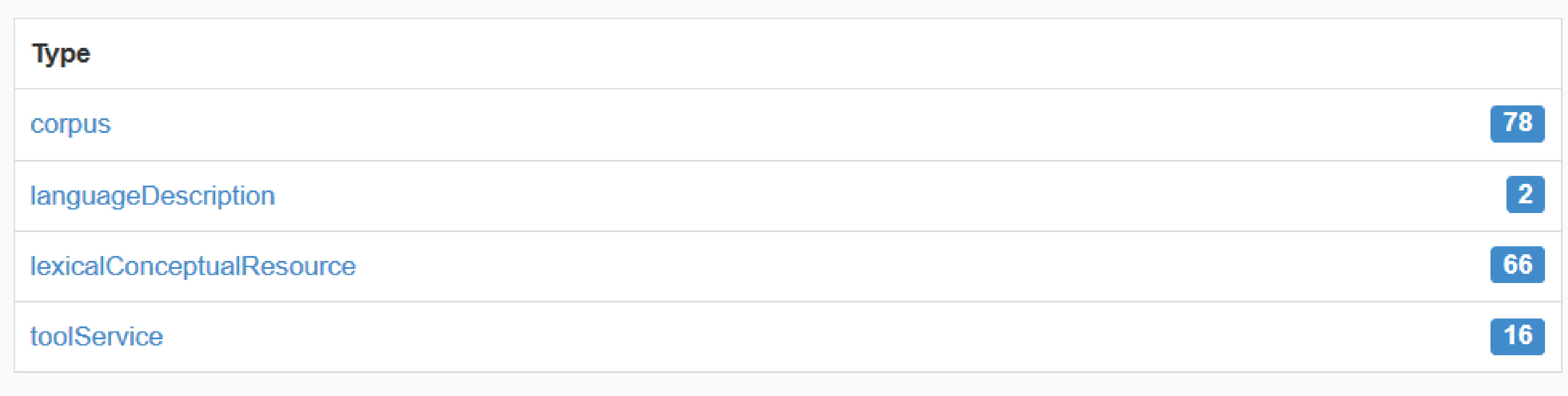
- Local authentication
- Clarin.eu website account 929 km
- Jožef Stefan Institute Slovenia 38 km
- ŠC Ptuj Slovenia Nearby
- School center Novo mesto Slovenia Nearby
- Gymnasium Piran Slovenia Nearby
- Gological Survey of Slovenia Slovenia Nearby
- IZUM Slovenia Nearby
- Gimnazija Bežigrad high school Slovenia Nearby

or search for a provider, such as Example University

Please help, I cannot find my provider

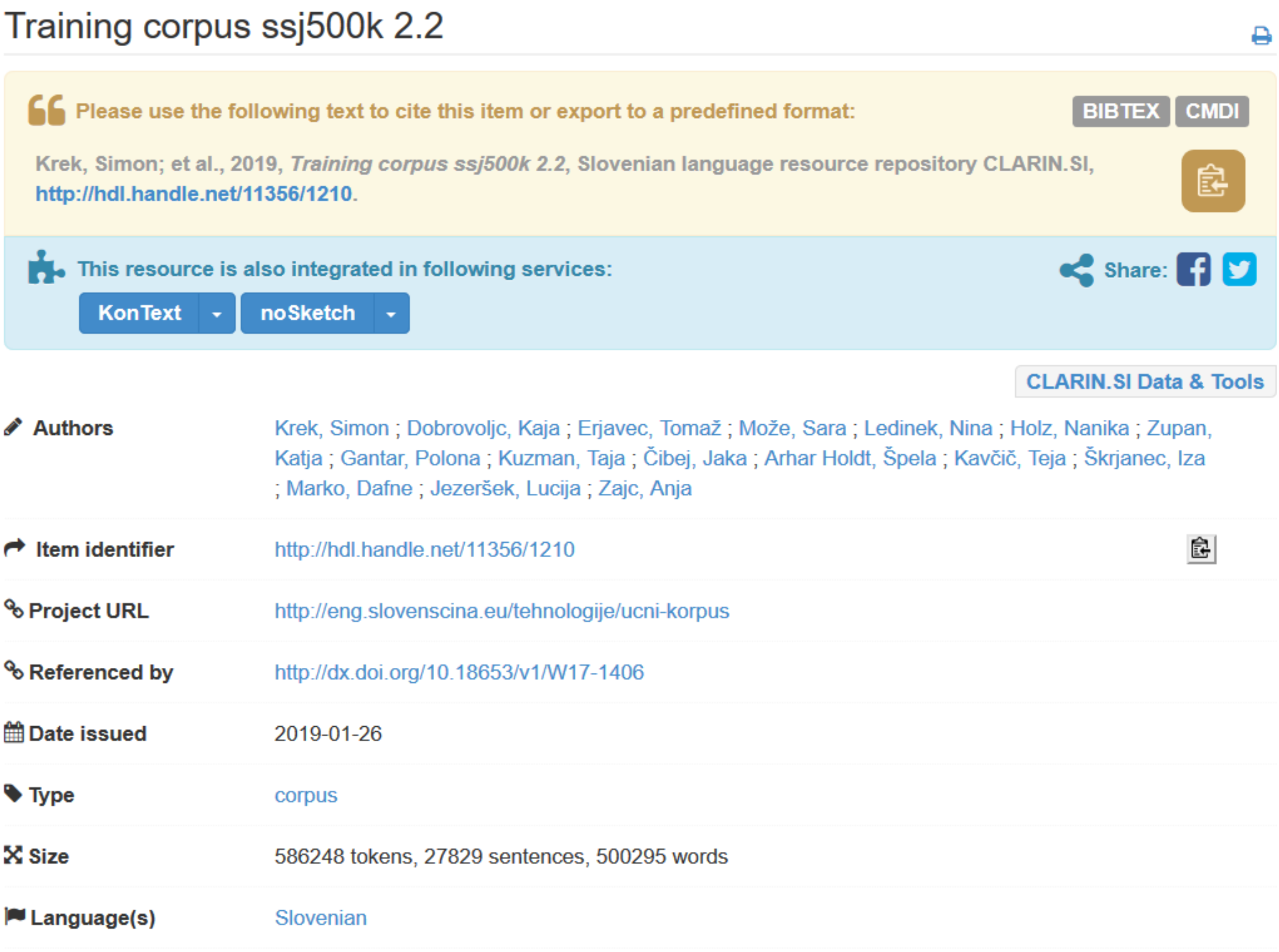
How to find interesting resources

- Browsing by language, type of resource, keywords, author, etc.
- Search (fuzzy matching)
- Advanced facet search
- Currently 163 items (without prior versions)



Type	
corpus	78
languageDescription	2
lexicalConceptualResource	66
toolService	16

Landing page of a repository item 1.



Training corpus ssj500k 2.2

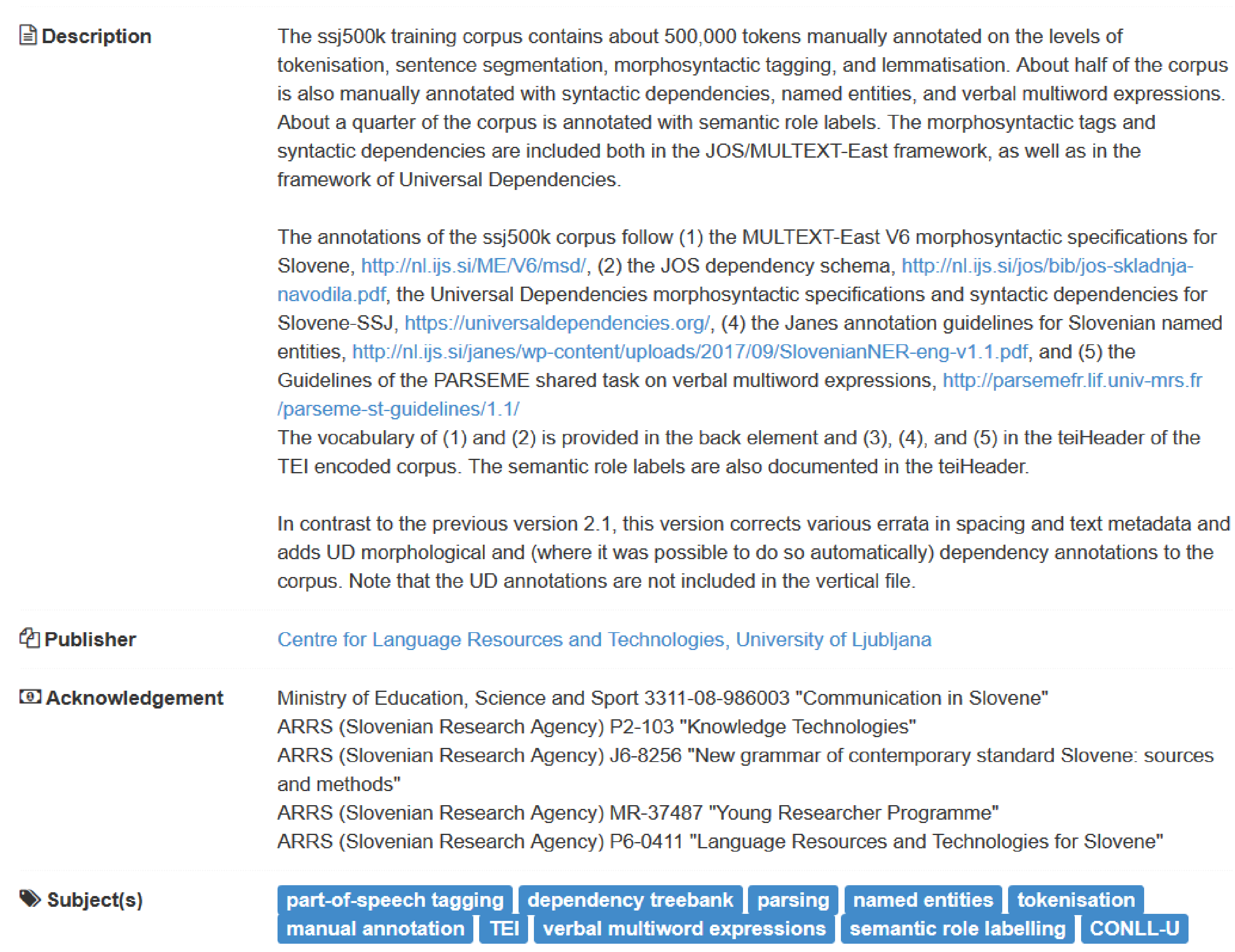
Please use the following text to cite this item or export to a predefined format:

Krek, Simon; et al., 2019, *Training corpus ssj500k 2.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1210.

This resource is also integrated in following services:

Authors	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/11356/1210
Project URL	http://eng.slovenscina.eu/tehnologije/ucni-korpus
Referenced by	http://dx.doi.org/10.18653/v1/W17-1406
Date issued	2019-01-26
Type	corpus
Size	586248 tokens, 27829 sentences, 500295 words
Language(s)	Slovenian

Landing page of a repository item 2



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 morphosyntactic tags and syntactic dependencies are included both in the JOS/MULTEXT-East framework, as well as in the framework of Universal Dependencies.

The annotations of the ssj500k corpus follow (1) the MULTEXT-East V6 morphosyntactic specifications for Slovene, http://nl.ijs.si/ME/V6/msd/, (2) the JOS dependency schema, http://nl.ijs.si/jos/bib/jos-skladnja-navodila.pdf, the Universal Dependencies morphosyntactic specifications and syntactic dependencies for Slovene-SSJ, https://universaldependencies.org/, (4) the Janes annotation guidelines for Slovenian named entities, http://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf, and (5) the Guidelines of the PARSEME shared task on verbal multiword expressions, http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/
The vocabulary of (1) and (2) is provided in the back element and (3), (4), and (5) in the teiHeader of the TEI encoded corpus. The semantic role labels are also documented in the teiHeader.

In contrast to the previous version 2.1, this version corrects various errata in spacing and text metadata and adds UD morphological and (where it was possible to do so automatically) dependency annotations to the corpus. Note that the UD annotations are not included in the vertical file.

Publisher

Centre for Language Resources and Technologies, University of Ljubljana

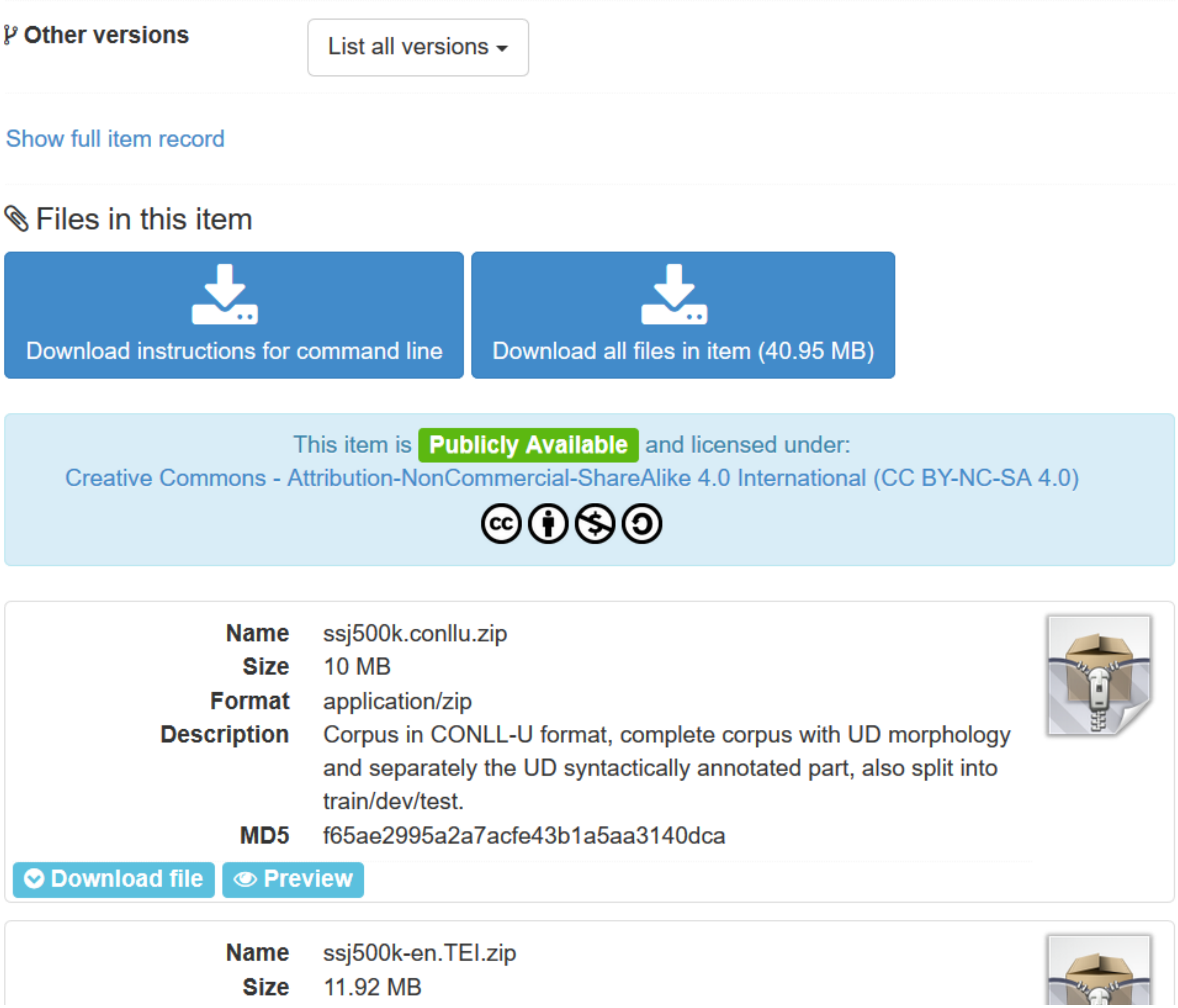
Acknowledgement

Ministry of Education, Science and Sport 3311-08-986003 "Communication in Slovene"
ARRS (Slovenian Research Agency) P2-103 "Knowledge Technologies"
ARRS (Slovenian Research Agency) J6-8256 "New grammar of contemporary standard Slovene: sources and methods"
ARRS (Slovenian Research Agency) MR-37487 "Young Researcher Programme"
ARRS (Slovenian Research Agency) P6-0411 "Language Resources and Technologies for Slovene"

Subject(s)

part-of-speech tagging | dependency treebank | parsing | named entities | tokenisation | manual annotation | TEI | verbal multiword expressions | semantic role labelling | CONLL-U

Landing page of a repository item 3


List all versions
Show full item record
Download instructions for command line
Download all files in item (40.95 MB)
Download file
Preview

Other versions

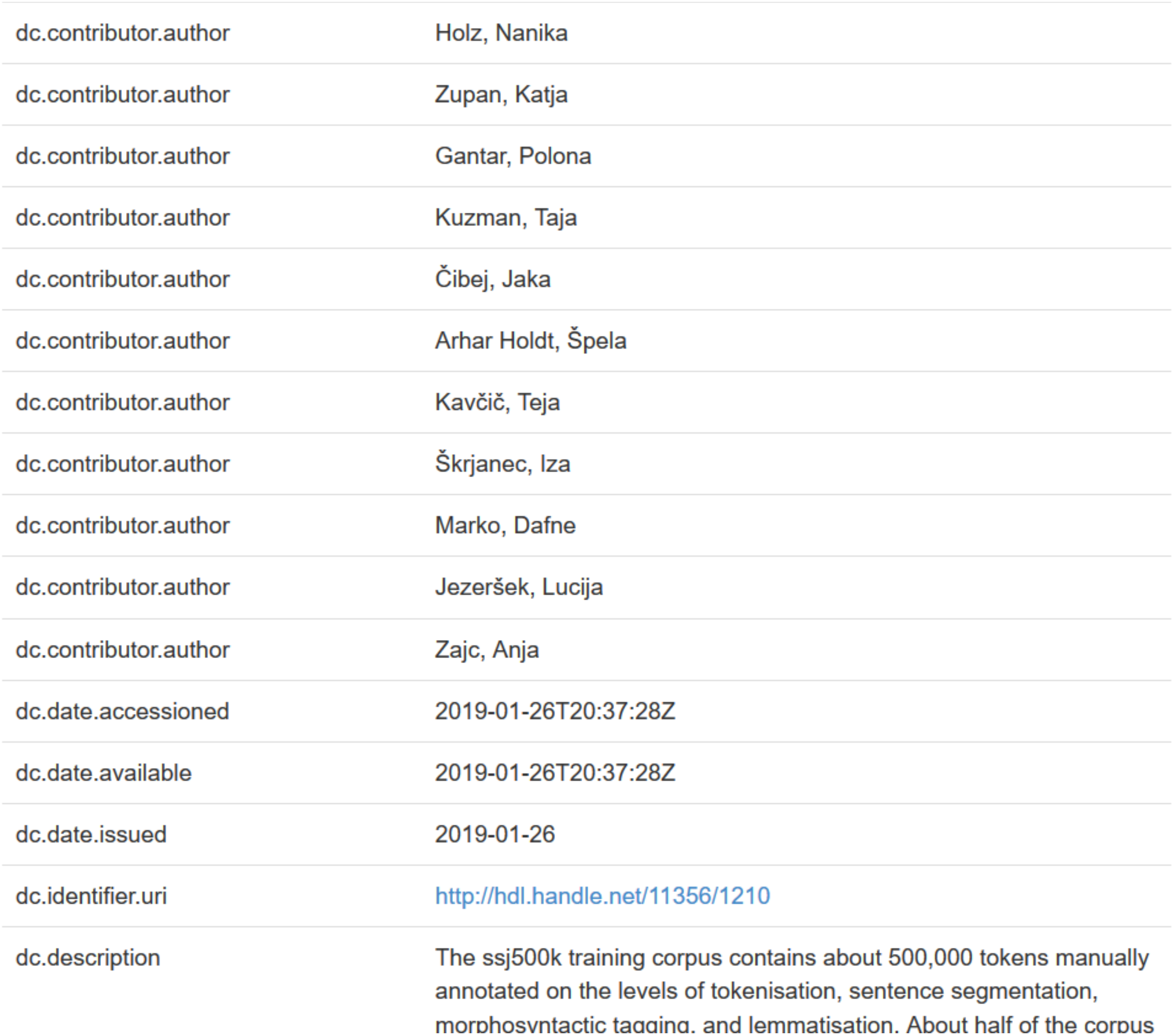
Files in this item

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

Name	ssj500k.conllu.zip
Size	10 MB
Format	application/zip
Description	Corpus in CONLL-U format, complete corpus with UD morphology and separately the UD syntactically annotated part, also split into train/dev/test.
MD5	f65ae2995a2a7acfe43b1a5aa3140dca

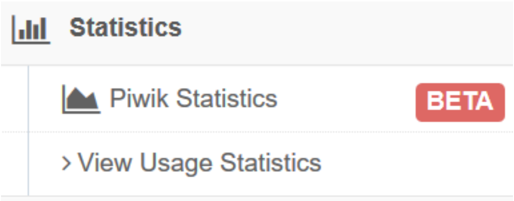
Name	ssj500k-en.TEI.zip
Size	11.92 MB

DC metadata



dc.contributor.author	Holz, Nanika
dc.contributor.author	Zupan, Katja
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	2019-01-26T20:37:28Z
dc.date.available	2019-01-26T20:37:28Z
dc.date.issued	2019-01-26
dc.identifier.uri	http://hdl.handle.net/11356/1210
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

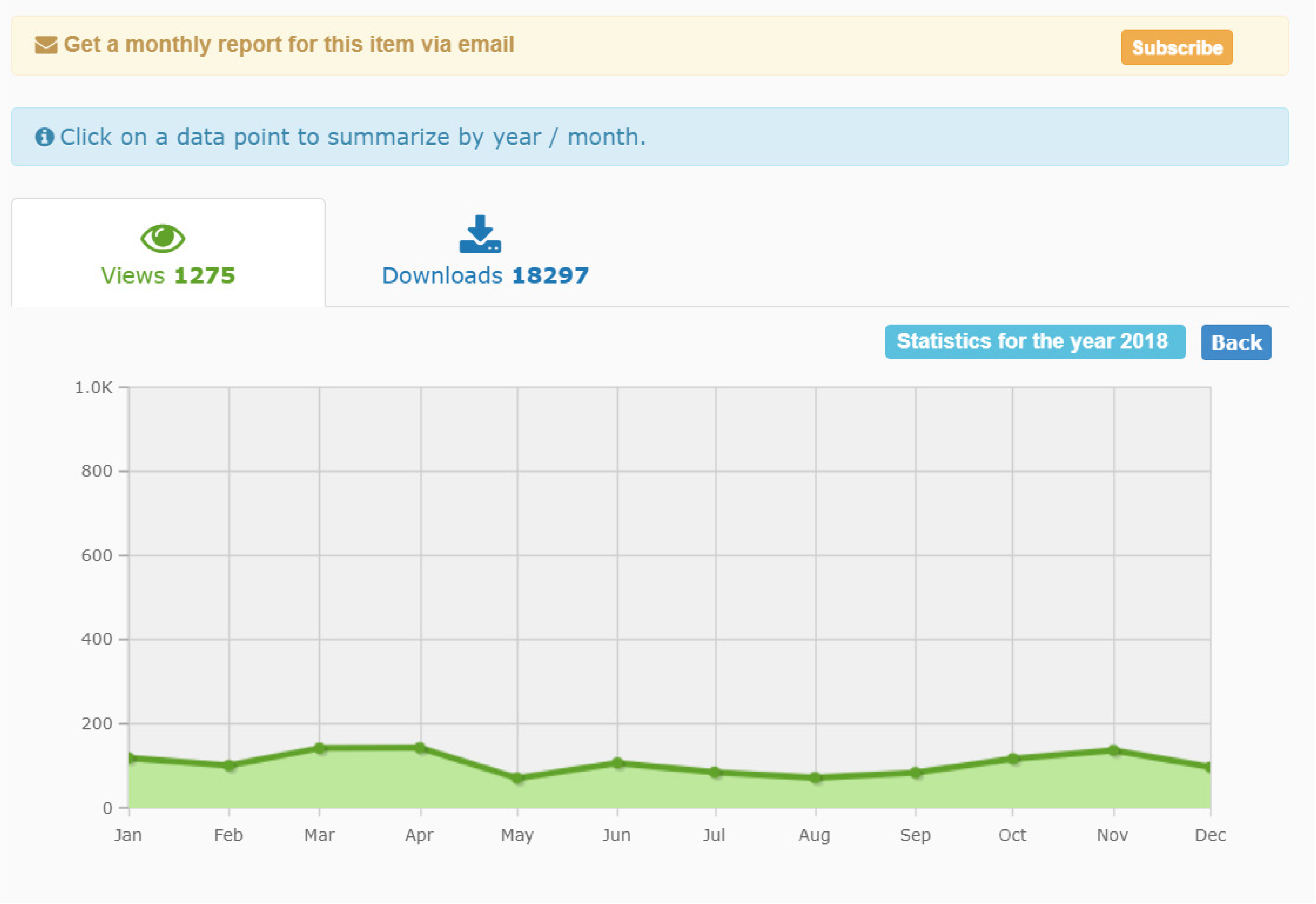
Piwik



Statistics

Piwik Statistics BETA

› View Usage Statistics



Get a monthly report for this item via email

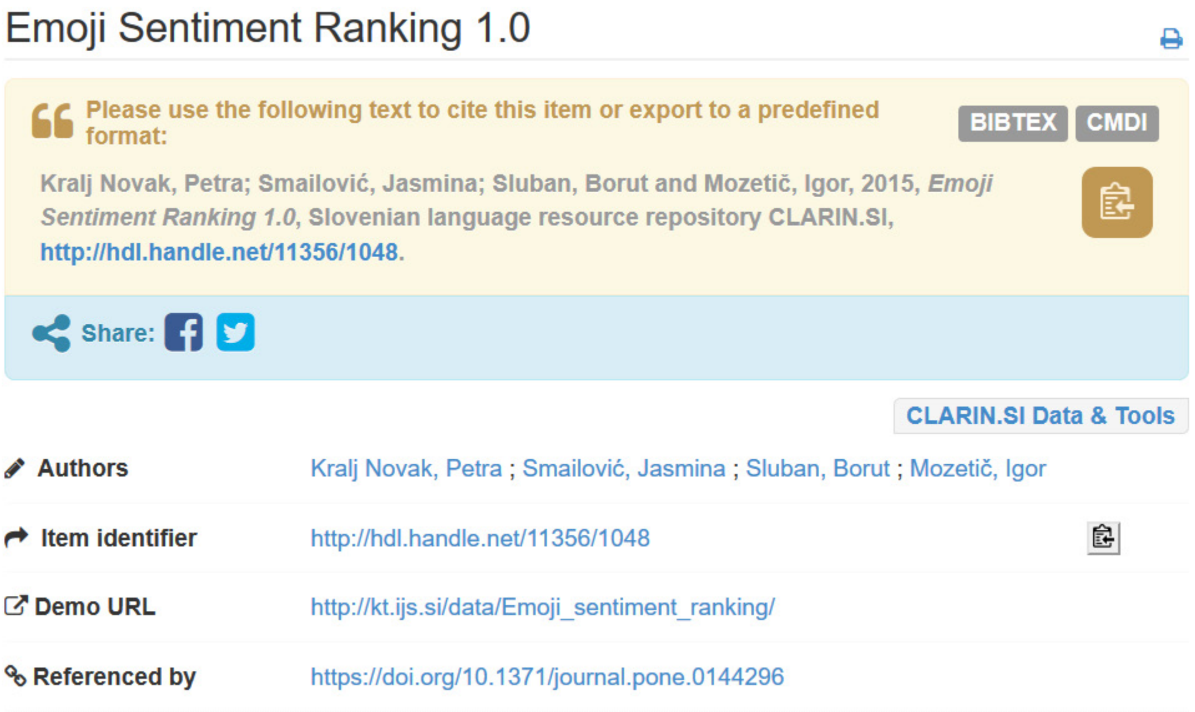
Click on a data point to summarize by year / month.

Views **1275**

Downloads **18297**

Repository: download

- Most entries available under Creative Commons licences
- Others: AAI login, agree to licence conditions: CLARIN.SI Licence ACA ID-BY-NC-INF-NORED
- Cite the resource!
- Always use the handle, not the URL!



Emoji Sentiment Ranking 1.0

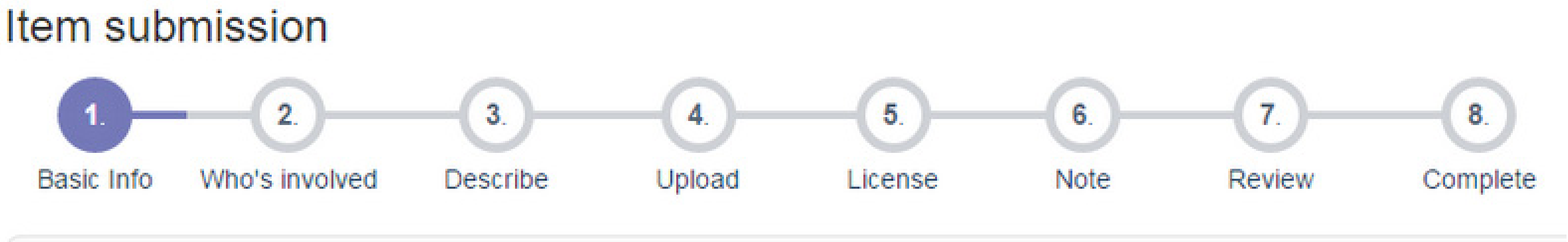
Please use the following text to cite this item or export to a predefined format:

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.

Authors	Kralj Novak, Petra ; Smailović, Jasmina ; Sluban, Borut ; Mozetič, Igor
Item identifier	http://hdl.handle.net/11356/1048
Demo URL	http://kt.ijs.si/data/Emoji_sentiment_ranking/
Referenced by	https://doi.org/10.1371/journal.pone.0144296

Repository: deposit

- AAI log-in
- Fairly simple workflow:



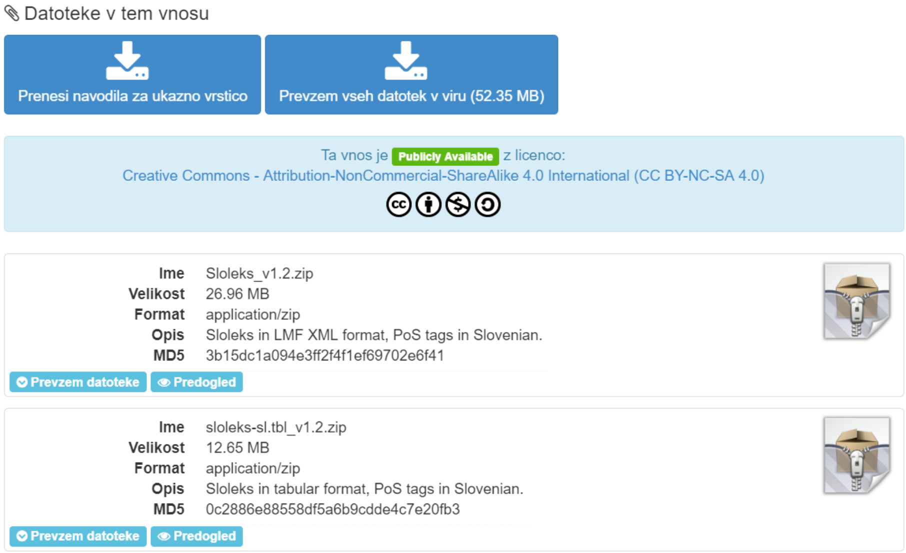
Item submission

1. Basic Info
2. Who's involved
3. Describe
4. Upload
5. License
6. Note
7. Review
8. Complete

- However: follow best practices, look at other entries!
- Choosing the licence
- Possibility to embargo the data
- Can enter a new version of a resource
- Option to share item editing
- The entry is checked by one of the CLARIN.SI editors

Data Formats

- Media files: wav, avi, jpg, png…
- Data with tabular structure: TSV, CSV, XML
- Hierarchical data: XML
  - using one of the „standard“ schemas, e.g. TEI, …, TRS, ELAN
  - using your own schema
  - in all cases: schema + documentation must be part of the entry
- + Derived formats + Documentation (can be PDF)
- Deposit as 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

What is in the repository?

www.slovenščina.eu

- Language models, training data, lexicons for tagging, lemmatisation, syntax of Slovene, Croatian, Serbian,
  for standard and non-standard language
- Various types of word embeddings
- Large annotated corpora of various text types
- Speech corpora
- Multilingual corpora
- Machine readable dictionaries
- Some software

MTE

АБУМАТРАН
ABUMATRAN

IMP

ReLDI

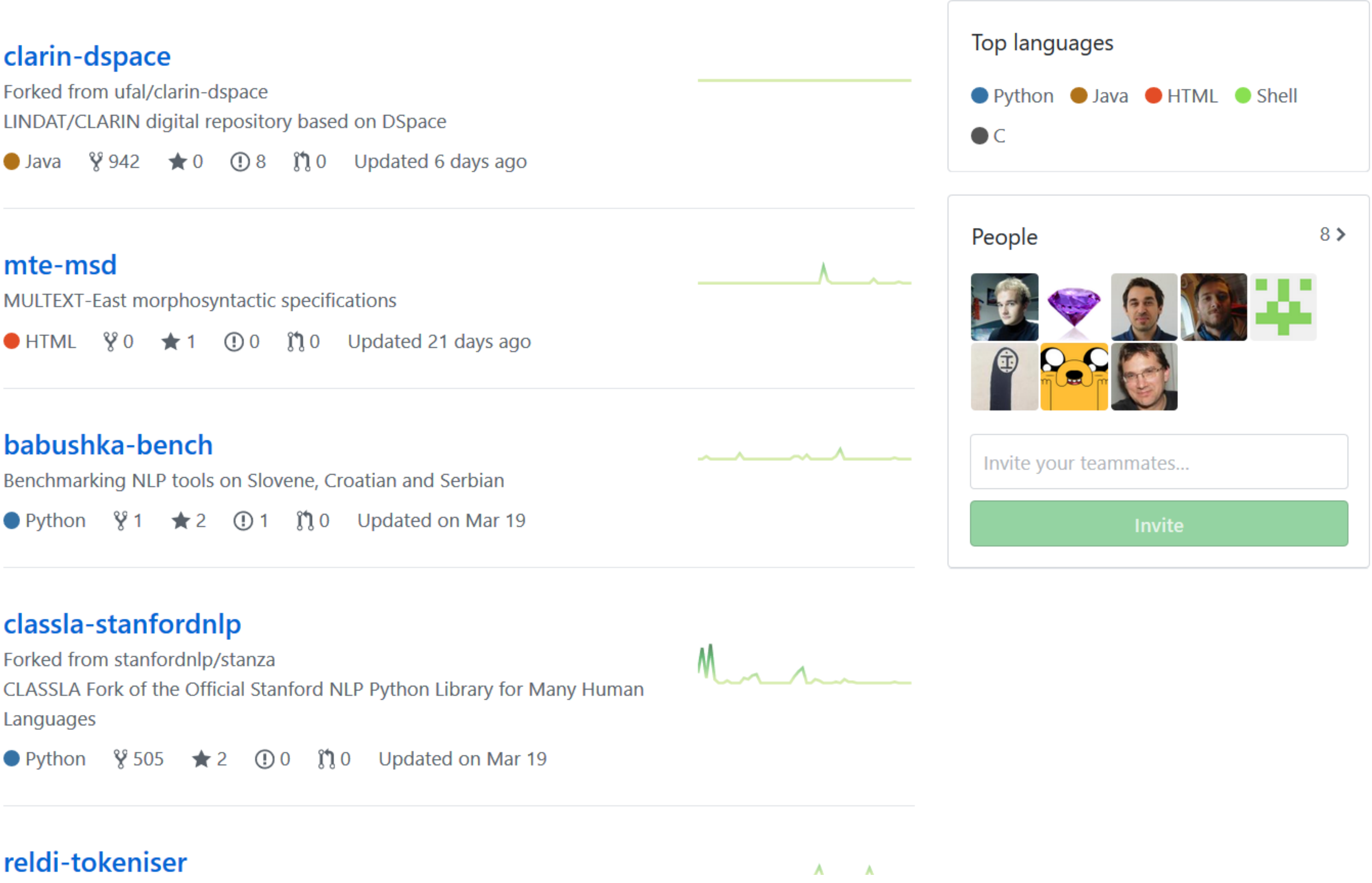
JANES

KAΣ
KORPUS
AKADEMSKE
SLOVENŠČINE

Tools

- “Language models, training data, lexicons
for tagging, lemmatisation, syntax of
Slovene, Croatian, Serbian,
for standard and non-standard language”
- tokenisers, morphosyntactic taggers for sl, hbs
- parsers (also UD-PIPE)
- Viewers and editors for linguistic annotation

CLARINSI@GitHub



clarin-dspace

Forked from ufal/clarin-dspace

LINDAT/CLARIN digital repository based on DSpace

● Java ⑂ 942 ★ 0 ⓘ 8 ⇅ 0 Updated 6 days ago

mte-msd

MULTEXT-East morphosyntactic specifications

● HTML ⑂ 0 ★ 1 ⓘ 0 ⇅ 0 Updated 21 days ago

babushka-bench

Benchmarking NLP tools on Slovene, Croatian and Serbian

● Python ⑂ 1 ★ 2 ⓘ 1 ⇅ 0 Updated on Mar 19

classla-stanfordnlp

Forked from stanfordnlp/stanza

CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages

● Python ⑂ 505 ★ 2 ⓘ 0 ⇅ 0 Updated on Mar 19

reldi-tokeniser

Top languages

● Python ● Java ● HTML ● Shell ● C

People 8 >

Invite your teammates...