The CLARIN.SI research infrastructure

Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute

FRI, 29th April, 2020

Overview of the lecture

1. Introduction
2. The CLARIN EU research infrastructure
3. The CLARIN.SI research infrastructure
4. CLARIN.SI services

I. Introduction

- Language technologies
  - main paradigm: supervised machine learning
  - programs are mostly language independent
  - need training (manually annotated) language resources
  - + test data
- Empirically supported linguistic investigations:
  - based on real (and, if possible, annotated) language data
- Annotated language resources are necessary for each language
- Where can we get such resources for Slovene (and other South-Slavic languages)?

Language resources

1. Corpora:
   - uniformly encoded and document collection of texts
   - explicit criteria for text selection
   - annotated (morphosyntax, lemmatisation, syntax, named entities, …)
   - reference/specialised; mono/multilingual; text/speech
2. Lexicons:
   - the vocabulary of a language
   - words / phrases
   - morphosyntax, syntax, semantics, translations, external and internal links
3. Models:
   - data that enables a program to annotate text in a certain language for a certain level of annotation
   - e.g. Stanford-NLP model for parsing of Slovene; Moses model translating Slovene to English

Resource reuse

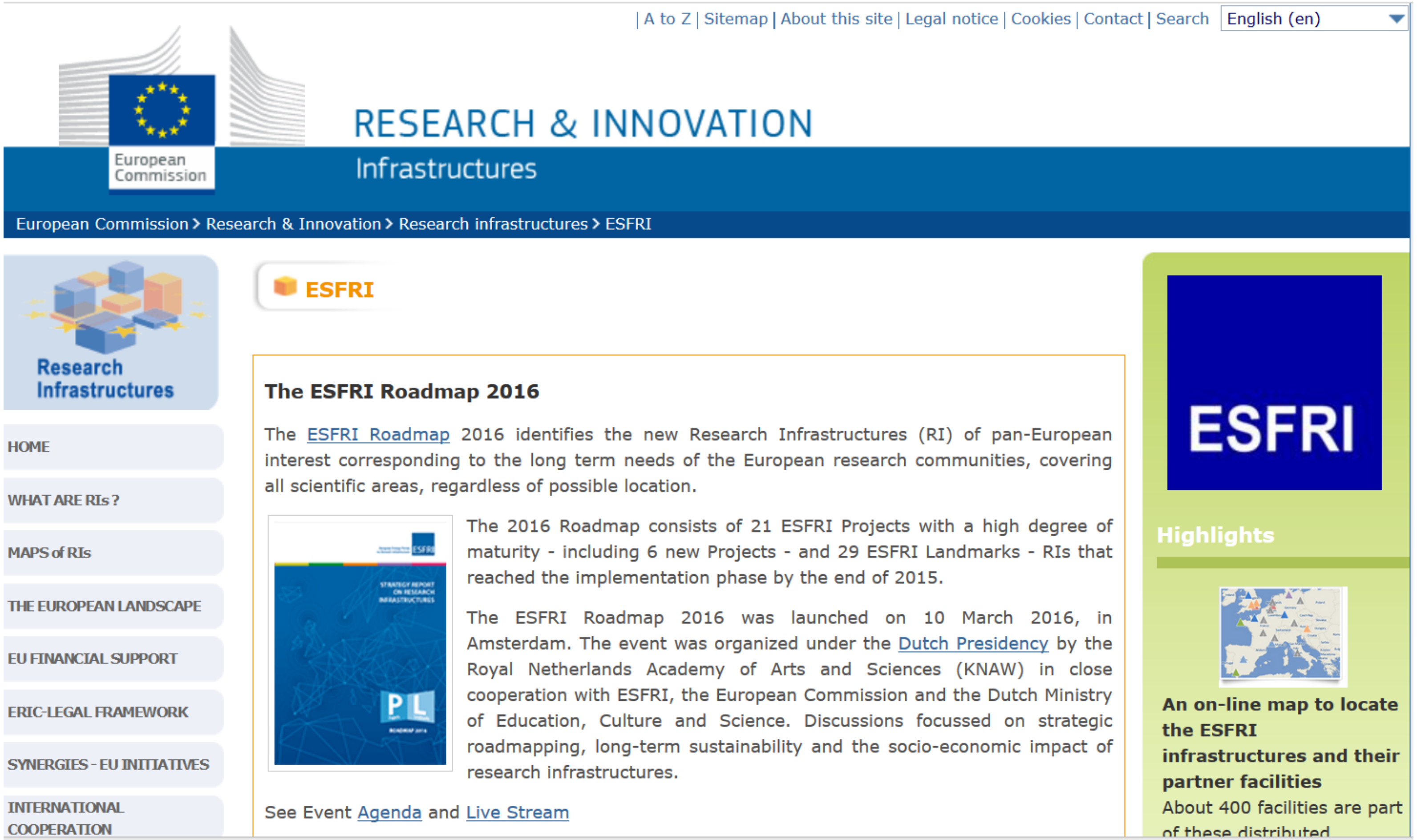
- Traditional approach:
  - develop language resources for each project separately
  - resources unavailable to other researchers
- Disadvantages:
  - the development of a language resource can be very costly: waste of time and money if it is done several times
  - later researchers cannot replicate or improve the initial results
  - supports the monopoly of institutions that produced the resources
  - the resources cannot be used to help in the development of products

Open access to the results of research projects

- No barriers to publications and data:
  - saves of time and money;
  - avoids repetition of work;
  - encourages cooperation;
  - makes the research process more transparent
  - stimulates innovation
- A very strong trend in EU (H2020) projects, also in Slovenia
- Problems in enabling open access to language resources:
  - copyright on texts
  - privacy protection (GDPR), including the right to be forgotten,
  - terms-of-use by owners of social media platforms (e.g. Twitter)

Research infrastructures

Research Infrastructures are facilities that provide resources and services for research communities to conduct research and foster innovation.



| A to Z | Sitemap | About this site | Legal notice | Cookies | Contact | Search
English (en)
European Commission
RESEARCH & INNOVATION
Infrastructures
European Commission > Research & Innovation > Research infrastructures > ESFRI
Research Infrastructures
HOME
WHAT ARE RIs ?
MAPS of RIs
THE EUROPEAN LANDSCAPE
EU FINANCIAL SUPPORT
ERIC-LEGAL FRAMEWORK
SYNERGIES - EU INITIATIVES
INTERNATIONAL COOPERATION
ESFRI
ESFRI
Highlights

The ESFRI Roadmap 2016

The ESFRI Roadmap 2016 identifies the new Research Infrastructures (RI) of pan-European interest corresponding to the long term needs of the European research communities, covering all scientific areas, regardless of possible location.

The 2016 Roadmap consists of 21 ESFRI Projects with a high degree of maturity - including 6 new Projects - and 29 ESFRI Landmarks - RIs that reached the implementation phase by the end of 2015.

The ESFRI Roadmap 2016 was launched on 10 March 2016, in Amsterdam. The event was organized under the Dutch Presidency by the Royal Netherlands Academy of Arts and Sciences (KNAW) in close cooperation with ESFRI, the European Commission and the Dutch Ministry of Education, Culture and Science. Discussions focussed on strategic roadmapping, long-term sustainability and the socio-economic impact of research infrastructures.

See Event Agenda and Live Stream

An on-line map to locate the ESFRI infrastructures and their partner facilities

About 400 facilities are part of these distributed

Research infrastructures

- Beginning, 2002: ESFRI (European Strategy Forum on Research Infrastructures),
- Roadmap: proposed 15 (2016: 21) RIs, some already established as ERICs (EU legal entity: European RI Consortium)
- Slovenia participates in 14 RI (e.g. CERN, ELEXIR)
- Humanities and Social Sciences:
  - DARIAH ERIC / DARIAH-SI: Digital Research Infrastructure for the Arts and Humanities
  - **CLARIN ERIC / CLARIN.SI**: Common Language Resources and Technology Infrastructure
  - Social Sciences: CESSDA / ADP, Arhiv družboslovnih podatkov

II. CLARIN ERIC

Common Language Resources and Technology Infrastructure

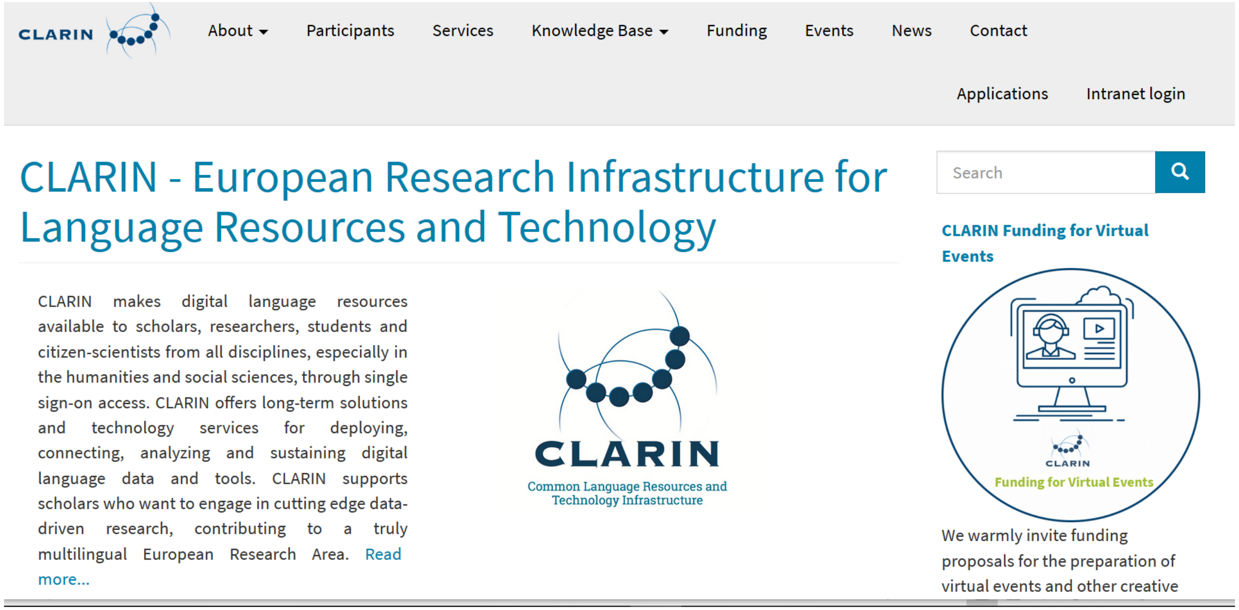
CLARIN

Common Language Resources and Technology Infrastructure

- Vision: digital language resources and technologies for all (European) languages are available for researchers in the humanities and social sciences
- Repository for long-term, extensive archiving and enabling access to language resources and technologies
- Contribution to preserving and supporting the European multilingual cultural heritage
- A collaborative paradigm in the compilation of language resources and the development of language tools, enabling re-use, experiment replicability and reproducibility

CLARIN

- Enable access to existing solutions in a unified infrastructure
- Consulting & teaching how to adapt tools and resources to specific research needs
- Legal, technical aspects of distribution
- Contribution to **standardisation of resources** and tools



CLARIN - European Research Infrastructure for Language Resources and Technology

CLARIN makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analyzing and sustaining digital language data and tools. CLARIN supports scholars who want to engage in cutting edge data-driven research, contributing to a truly multilingual European Research Area. Read more...

CLARIN Funding for Virtual Events

We warmly invite funding proposals for the preparation of virtual events and other creative

CLARIN ERIC

- 21 member states + 4 observers
- Based in the Netherlands:
  director, support staff, strong DH / CL community
- Committees: BoD, NCF, SCTC, …
- Aggregators: Virtual Language Observatory
- Most work is done by the national consortia
- Annual conference:
  - authors of accepted paper go for free
  - session for PhD students
  - book of abstracts (post-conference papers), posters, bazaar, invited talks etc.

III. CLARIN.SI

CLARIN.SI

- CLARIN Slovenia, start of work in 2014
- Organised as a consortium of (currently) 11 partners:
  - 4 universities: Ljubljana, Maribor, Nova Gorica, Primorska
  - 4 research institutes: ZRC SAZU, IJS, INZ, Trojina
  - 2 companies: Amebis, Alpineon
  - 1 society: Slovenian society for language technologies, SDJT
- Headquarters at IJS:
  - E8: Dept. for Knowledge Technologies
  - E3: Laboratory for Artificial Intelligence
  - CMI: Networking Infrastructure Centre<

CLARIN.SI

- Repository
  - Long term archiving of language resources (and tools)
  - Also, for software and manually annotated datasets: CLARINSI GitHub virtual organisation & http://gitlab.clarin.si
- Web services:
  - 2 concordancers (corpus analysis)
  - automatic annotation
  - WebAnno platform for manual annotation (e.g. training sets)
- Support for events:
  - Conference „Language Technologies and digital humanities“ (1998, …, 2016, 2018, 2020)
  - JOTA lectures “Jezikovnotehnološki abonma”: VideoLectures
  - XVIII EURALEX International Congress, Ljubljana, 2018
  - 22nd Intl. Conf. on Text Speech and Dialogue, Ljubljana, 2019
- Support for development and archiving language resources and tools
  - support for resource update for archiving in the repository (cca 500 EUR)
  - larger projects for development: 2018: 8, 2019: 7 projects (cca 6,000 EUR)

CLASSLA

CLARIN
K CENTRE

- CLARIN certified knowledge centre for Processing of South Slavic languages
- CLARIN.SI + Bulgarian CLARIN
- FAQ on processing Slovenian, Croatian, Serbian, Bulgarian
- CLASSLA automatic annotation web service
- CLARIN.SI repository offers the most resources for Croatian and Serbian
- CLARINSI@GitHub offers many tools to process Slovenian, Croatian and Serbian (HBS)

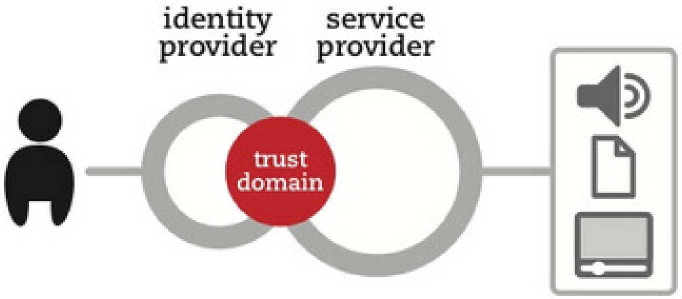
CLARIN.SI Cooperation

- CLARIN: National coordinators forum, Working Groups on Standards, Legal Issues, User Involvement, Technical Centres
- DARIAH-SI (INZ): joint development of corpora: digital library + linguistically analysed corpus (e.g. siParl)
- ADP/CESSDA (FDV): RDA Node Slovenia

IV. CLARIN.SI Services

- AAI Log-in
- Concordancers
- WebAnno
- ReLDI annotation
- Repository and Git (next lecture)

Log-in

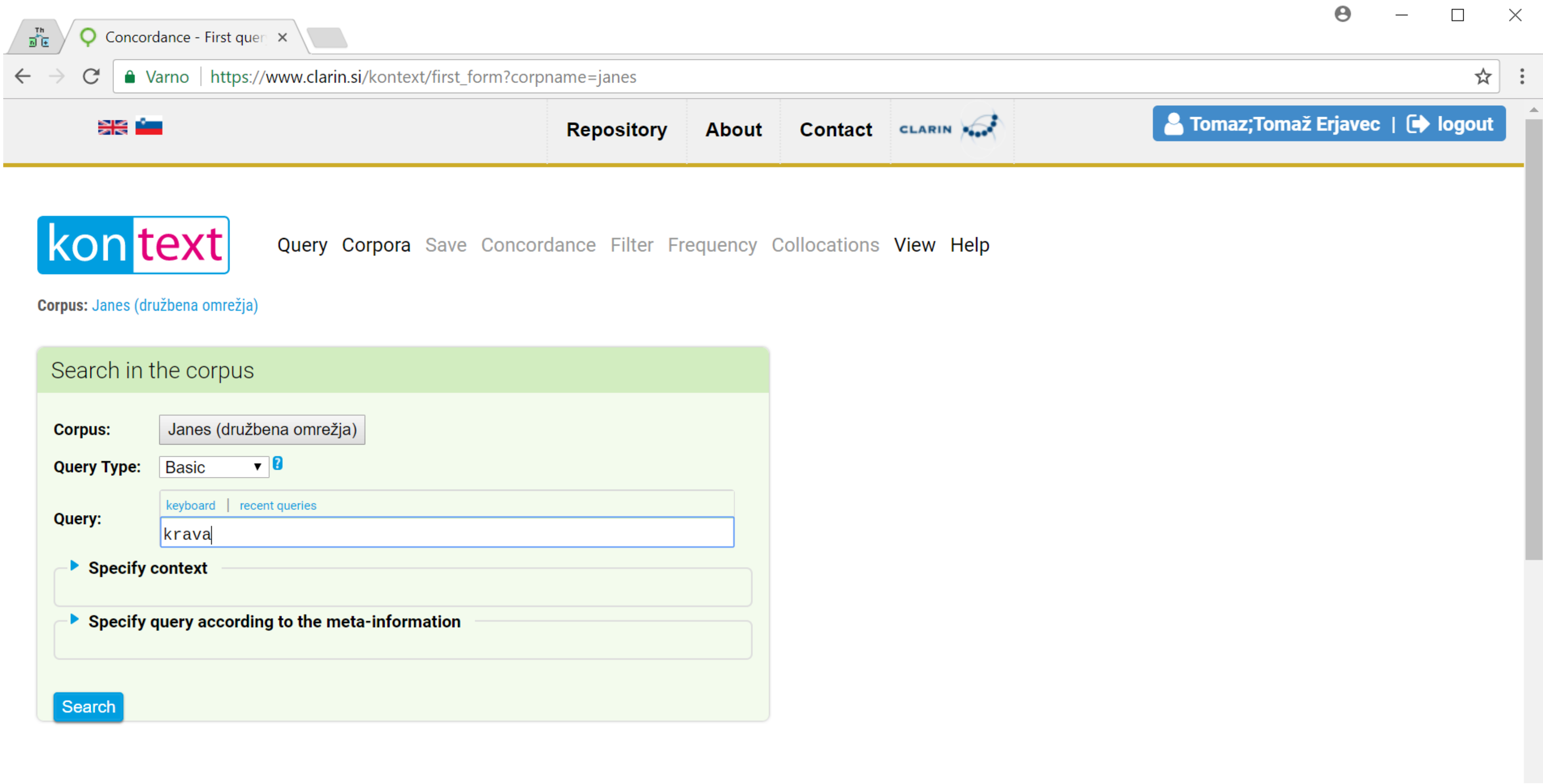


- Infrastructure for authentication and authorisation (AAI)
- Single Sign-On: separation between the Identity Provider and Service Provider
- As opposed to standard web login we here know the identity of the user
- Identity provider federations: EduGain
- Slovene users can, via EduGain, access most CLARIN services in Europe

Concordancers

- KonText + noSketch Engine
- both use the same back-end: Manatee
- support large corpora (> billion words)
- corpora can have rich annotations:
  - structures (text, speech, sentence, etc.)
  - meta-data (publication date, text type, author name, text standardness, etc.)
  - token attributes (PoS tag, lemma, normalised form, etc.)
- powerful query language: CQL
- various types of analysis and output
- RESTFUL interface: usable via API
- CLARIN.SI concordancers offer ~100 corpora

KonText (CLARIN-CZ)



Concordance - First quer
Varno
https://www.clarin.si/kontext/first_form?corpname=janes
Repository
About
Contact
CLARIN
Tomaz;Tomaž Erjavec | logout
kon text
Query
Corpora
Save
Concordance
Filter
Frequency
Collocations
View
Help
keyboard | recent queries
Specify context
Specify query according to the meta-information
Search

Corpus: Janes (družbena omrežja)

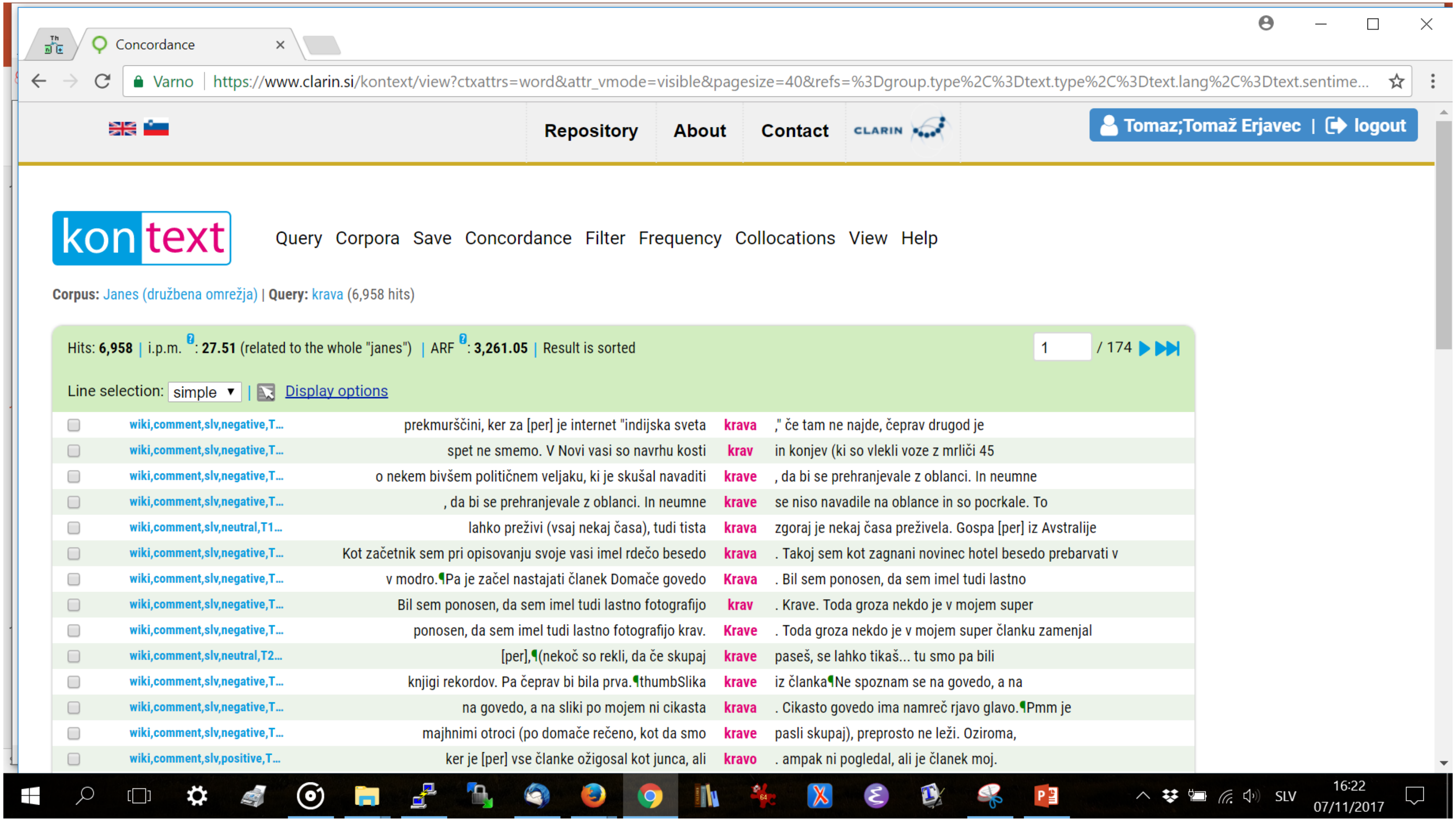
Search in the corpus

Corpus: Janes (družbena omrežja)

Query Type: Basic

Query: krava

Concordances


Concordance
Varno
https://www.clarin.si/kontext/view?ctxattrs=word&attr_vmode=visible&pagesize=40&refs=%3Dgroup.type%2C%3Dtext.type%2C%3Dtext.lang%2C%3Dtext.sentime...
Repository
About
Contact
CLARIN
Tomaz;Tomaž Erjavec | logout
kontext
Query
Corpora
Save
Concordance
Filter
Frequency
Collocations
View
Help
SLV
16:22
07/11/2017

Corpus: Janes (družbena omrežja) | **Query:** krava (6,958 hits)

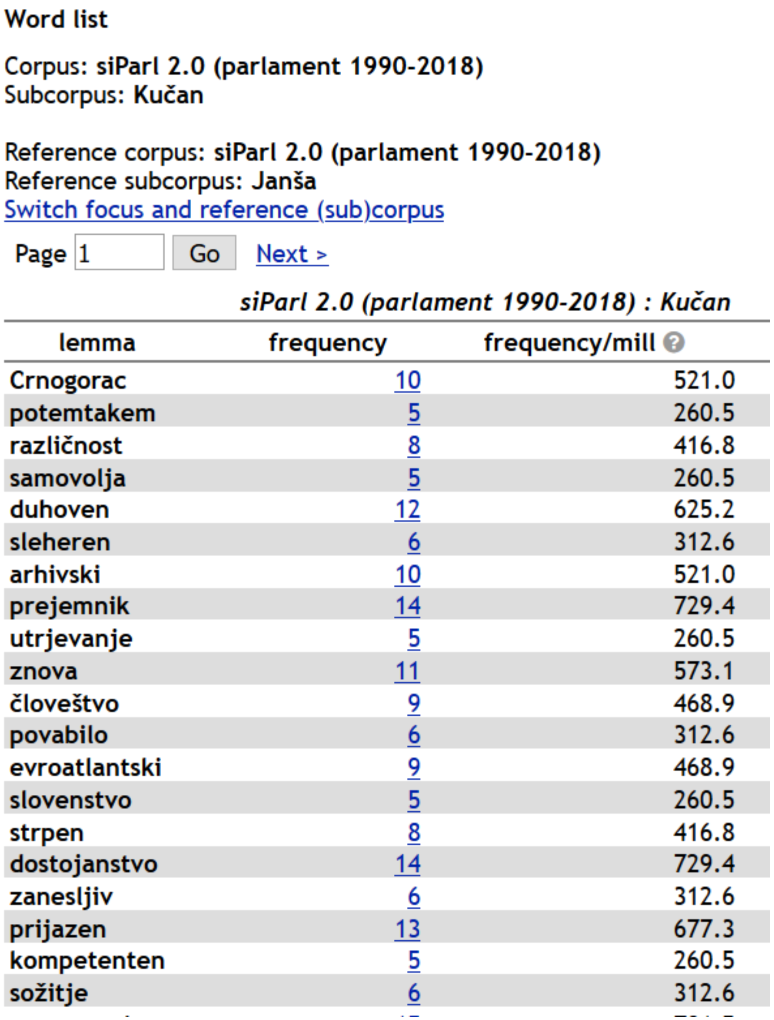
Hits: **6,958** | i.p.m. ?: **27.51** (related to the whole "janes") | ARF ?: **3,261.05** | Result is sorted

1 / 174

Line selection: simple ▾ | Display options

□	wiki,comment,slv,negative,T...	prekmurščini, ker za [per] je internet "indijska sveta	krava	," če tam ne najde, čeprav drugod je
□	wiki,comment,slv,negative,T...	spet ne smemo. V Novi vasi so navrhu kosti	krav	in konjev (ki so vlekli voze z mrliči 45
□	wiki,comment,slv,negative,T...	o nekem bivšem političnem veljaku, ki je skušal navaditi	krave	, da bi se prehranjevale z oblanci. In neumne
□	wiki,comment,slv,negative,T...	, da bi se prehranjevale z oblanci. In neumne	krave	se niso navadile na oblance in so pocrkale. To
□	wiki,comment,slv,neutral,T1...	lahko preživi (vsaj nekaj časa), tudi tista	krava	zgoraj je nekaj časa preživela. Gospa [per] iz Avstralije
□	wiki,comment,slv,negative,T...	Kot začetnik sem pri opisovanju svoje vasi imel rdečo besedo	krava	. Takoj sem kot zagnani novinec hotel besedo prebarvati v
□	wiki,comment,slv,negative,T...	v modro.¶Pa je začel nastajati članek Domače govedo	Krava	. Bil sem ponosen, da sem imel tudi lastno
□	wiki,comment,slv,negative,T...	Bil sem ponosen, da sem imel tudi lastno fotografijo	krav	. Krave. Toda groza nekdo je v mojem super
□	wiki,comment,slv,negative,T...	ponosen, da sem imel tudi lastno fotografijo krav.	Krave	. Toda groza nekdo je v mojem super članku zamenjal
□	wiki,comment,slv,neutral,T2...	[per],¶(nekoč so rekli, da če skupaj	krave	paseš, se lahko tikaš... tu smo pa bili
□	wiki,comment,slv,negative,T...	knjigi rekordov. Pa čeprav bi bila prva.¶thumbSlika	krave	iz članka¶Ne spoznam se na govedo, a na
□	wiki,comment,slv,negative,T...	na govedo, a na sliki po mojem ni cikasta	krava	. Cikasto govedo ima namreč rjavo glavo.¶Pmm je
□	wiki,comment,slv,negative,T...	majhnimi otroci (po domače rečeno, kot da smo	krave	pasli skupaj), preprosto ne leži. Oziroma,
□	wiki,comment,slv,positive,T...	ker je [per] vse članke ožigosal kot junca, ali	kravo	. ampak ni pogledal, ali je članek moj.

Kučan vs. Janša



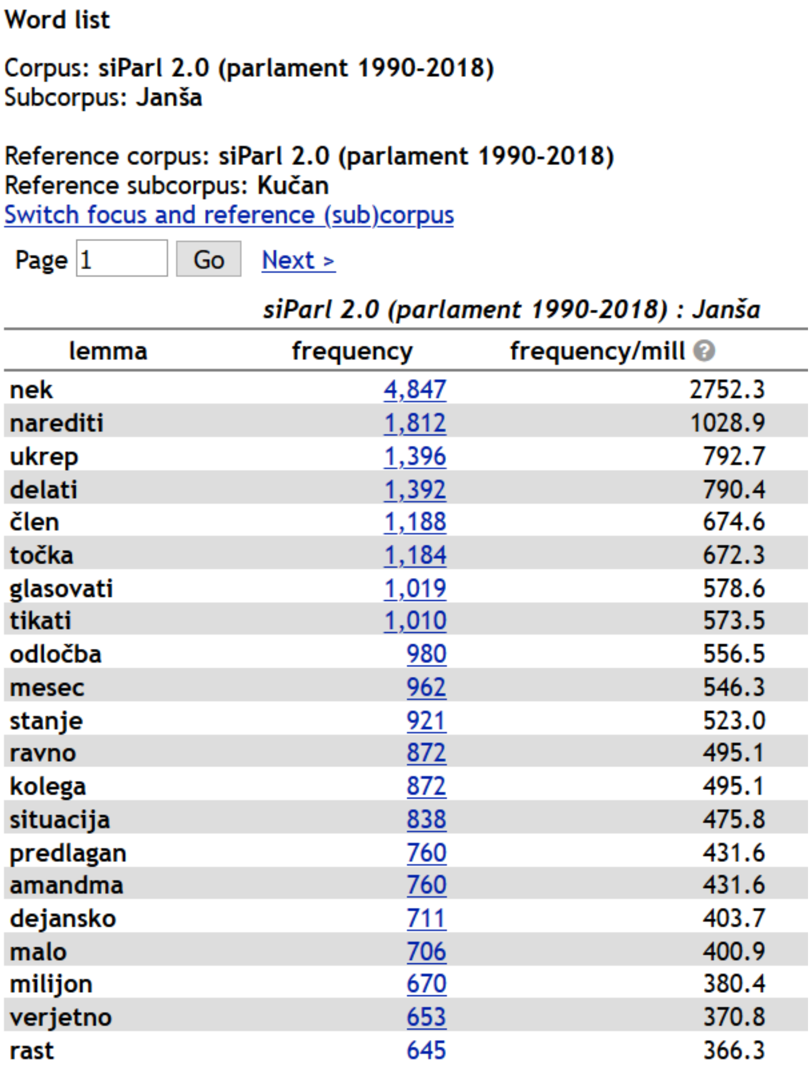
Word list

Corpus: **siParl 2.0 (parlament 1990-2018)**
Subcorpus: **Kučan**

Reference corpus: **siParl 2.0 (parlament 1990-2018)**
Reference subcorpus: **Janša**
Switch focus and reference (sub)corpus

siParl 2.0 (parlament 1990-2018) : Kučan

lemma	frequency	frequency/mill
Crnogorac	10	521.0
potemtakem	5	260.5
različnost	8	416.8
samovolja	5	260.5
duhoven	12	625.2
sleheren	6	312.6
arhivski	10	521.0
prejemnik	14	729.4
utrjevanje	5	260.5
znova	11	573.1
človeštvo	9	468.9
povabilo	6	312.6
evroatlantski	9	468.9
slovenstvo	5	260.5
strpen	8	416.8
dostojanstvo	14	729.4
zanesljiv	6	312.6
prijazen	13	677.3
kompetenten	5	260.5
sožitje	6	312.6



Word list

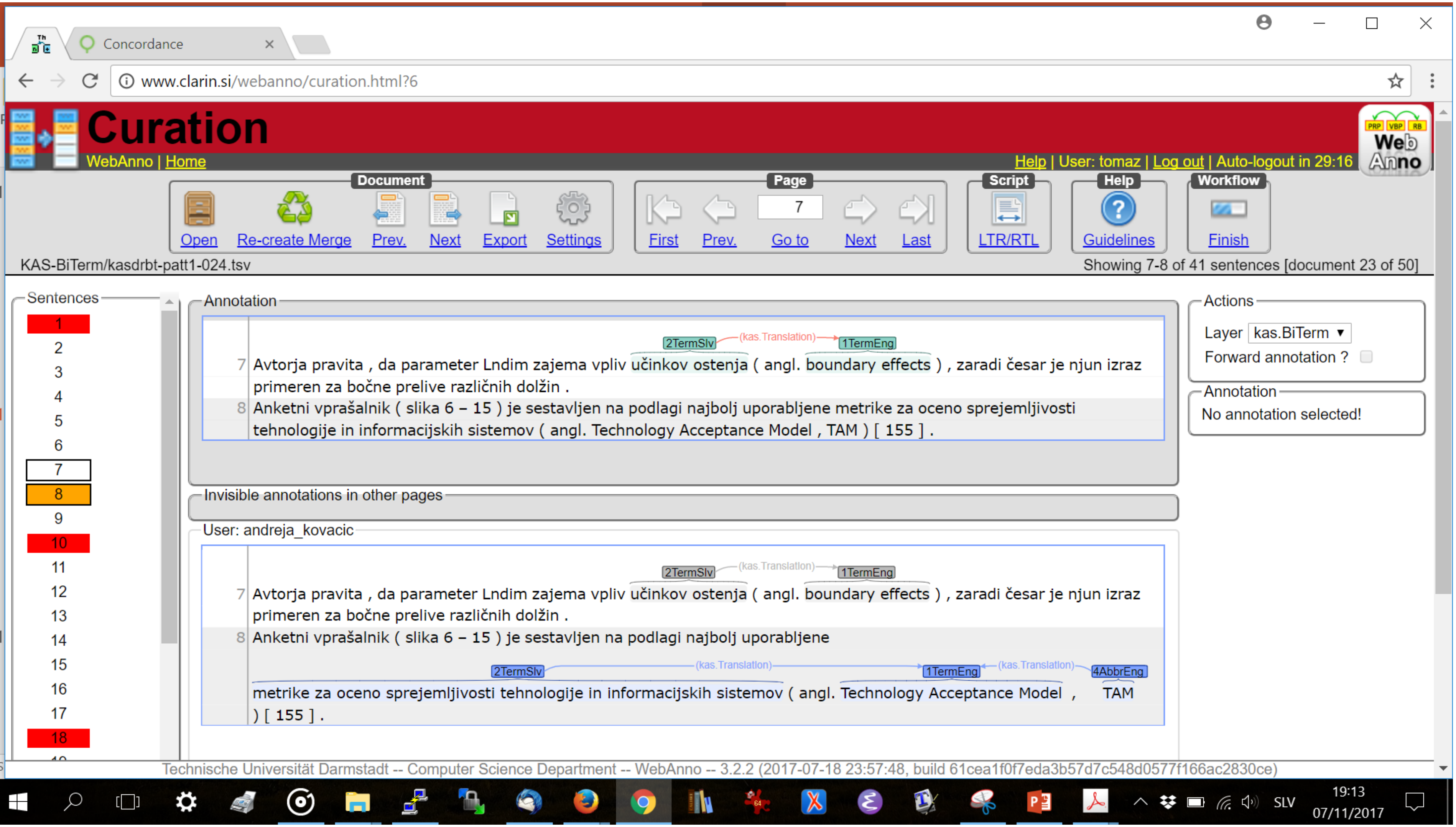
Corpus: **siParl 2.0 (parlament 1990-2018)**
Subcorpus: **Janša**

Reference corpus: **siParl 2.0 (parlament 1990-2018)**
Reference subcorpus: **Kučan**
Switch focus and reference (sub)corpus

siParl 2.0 (parlament 1990-2018) : Janša

lemma	frequency	frequency/mill
nek	4,847	2752.3
narediti	1,812	1028.9
ukrep	1,396	792.7
delati	1,392	790.4
član	1,188	674.6
točka	1,184	672.3
glasovati	1,019	578.6
tikati	1,010	573.5
odločba	980	556.5
mesec	962	546.3
stanje	921	523.0
ravno	872	495.1
kolega	872	495.1
situacija	838	475.8
predlagan	760	431.6
amandma	760	431.6
dejansko	711	403.7
malo	706	400.9
milijon	670	380.4
verjetno	653	370.8
rast	645	366.3

WebAnno (CLARIN-DE)



KAS-BiTerm/kasdrbt-patt1-024.tsv

Showing 7-8 of 41 sentences [document 23 of 50]

Annotation

7 Avtorja pravita , da parameter Lndim zajema vpliv učinkov ostenja (angl. boundary effects) , zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .

8 Anketni vprašalnik (slika 6 – 15) je sestavljen na podlagi najbolj uporabljene metrike za oceno sprejemljivosti tehnologije in informacijskih sistemov (angl. Technology Acceptance Model , TAM) [155] .

User: andreja_kovacic

7 Avtorja pravita , da parameter Lndim zajema vpliv učinkov ostenja (angl. boundary effects) , zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .

8 Anketni vprašalnik (slika 6 – 15) je sestavljen na podlagi najbolj uporabljene metrike za oceno sprejemljivosti tehnologije in informacijskih sistemov (angl. Technology Acceptance Model , TAM) [155] .

Actions

Layer: kas.BiTerm

Forward annotation ? □

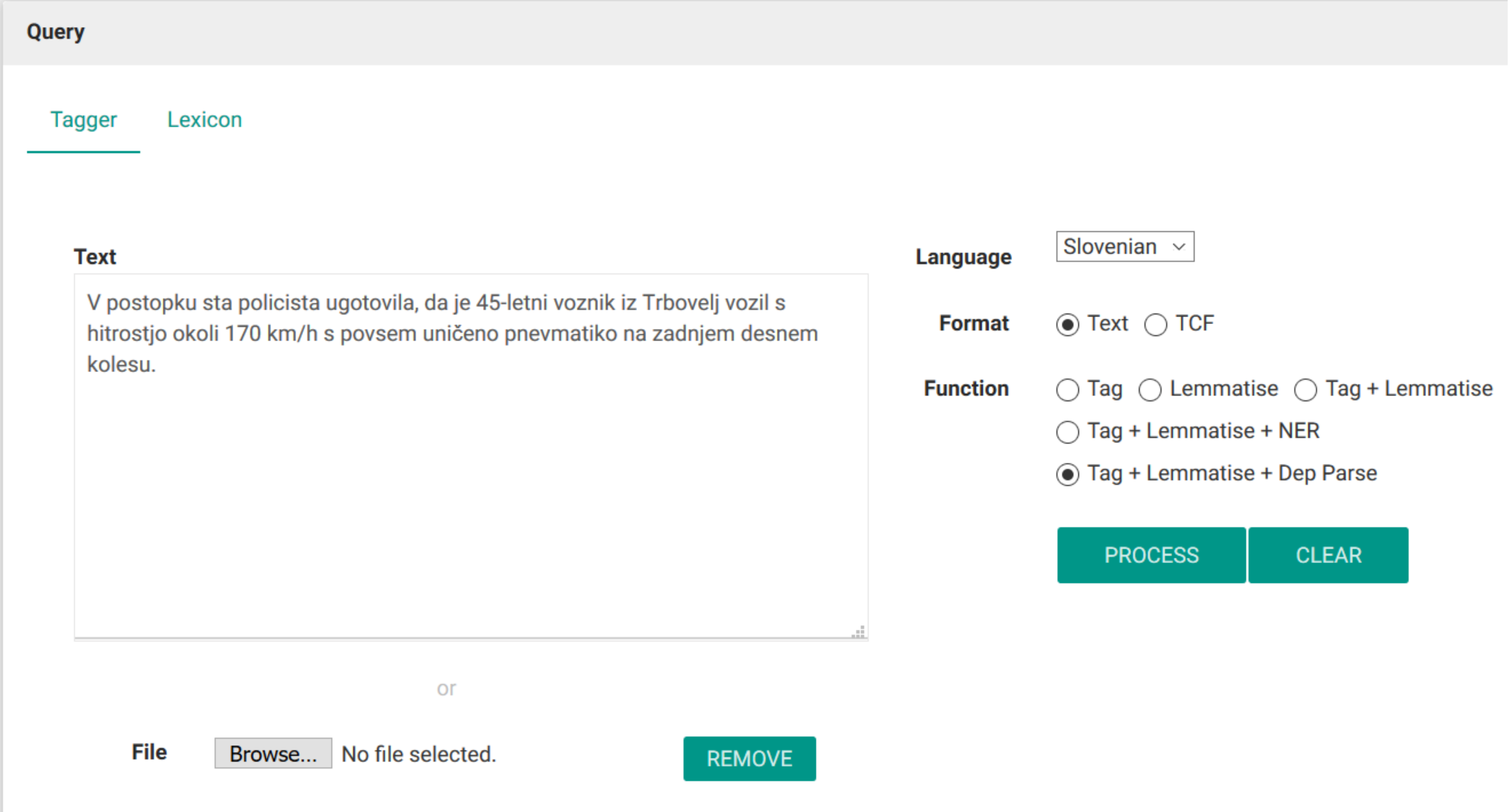
Annotation

No annotation selected!

Technische Universität Darmstadt -- Computer Science Department -- WebAnno -- 3.2.2 (2017-07-18 23:57:48, build 61cea1f0f7eda3b57d7c548d0577f166ac2830ce)

ReLDI automatic text annotation

- Web form
- API



Query

Text

V postopku sta policista ugotovila, da je 45-letni voznik iz Trbovelj vozil s hitrostjo okoli 170 km/h s povsem uničeno pnevmatiko na zadnjem desnem kolesu.

Language Slovenian

Format ● Text ○ TCF

Function ○ Tag ○ Lemmatise ○ Tag + Lemmatise ○ Tag + Lemmatise + NER ● Tag + Lemmatise + Dep Parse

File Browse... No file selected.

Result

	Surface	Tags	Lemma	Dep parse - gov / func
1.	V	Sl	v	2 / case
2.	postopku	Ncmsl	postopek	3 / nmod
3.	sta	Va-r3d-n	biti	0 / root
4.	policista	Ncmdn	policist	3 / dobj
5.	ugotovila	Vmep-dm	ugotoviti	4 / acl
6.	,	Z	,	5 / punct
7.	da	Cs	da	13 / mark
8.	je	Va-r3s-n	biti	13 / aux
9.	45-letni	Agpmsny	45-leten	10 / amod
10.	voznik	Ncmsn	voznik	13 / nsubj
11.	iz	Sg	iz	12 / case
12.	Trbovelj	Npfpg	Trbovlje	10 / nmod

V. Conclusions

- The purpose of CLARIN(.SI) is to support research that need access to language data
    - Digital humanities and social sciences
    - Language Technologies (~ Computational Linguistics)
    - All other fields where language is important
- Open access to resources, tools and services
- Where authentication is needed, AAI is used
- CLARIN(.SI) financial support:
    - Organising various types of events
    - Work on specific topics incl. outreach
    - Development or modification of resources
    - Attendance at CLARIN conferences

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:

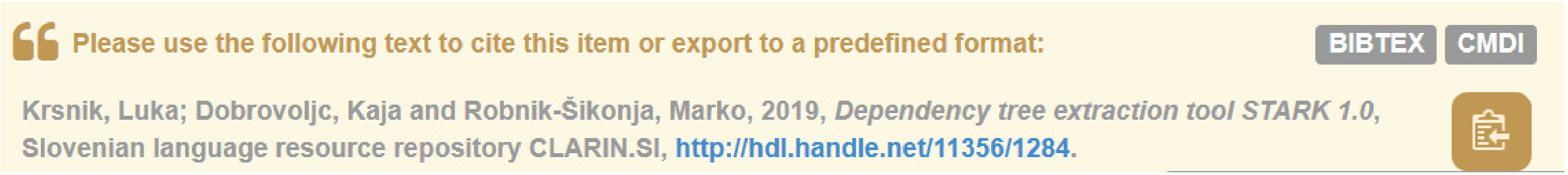
OPEN ARCHIVES

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

OpenAIRE

Permanent identifiers

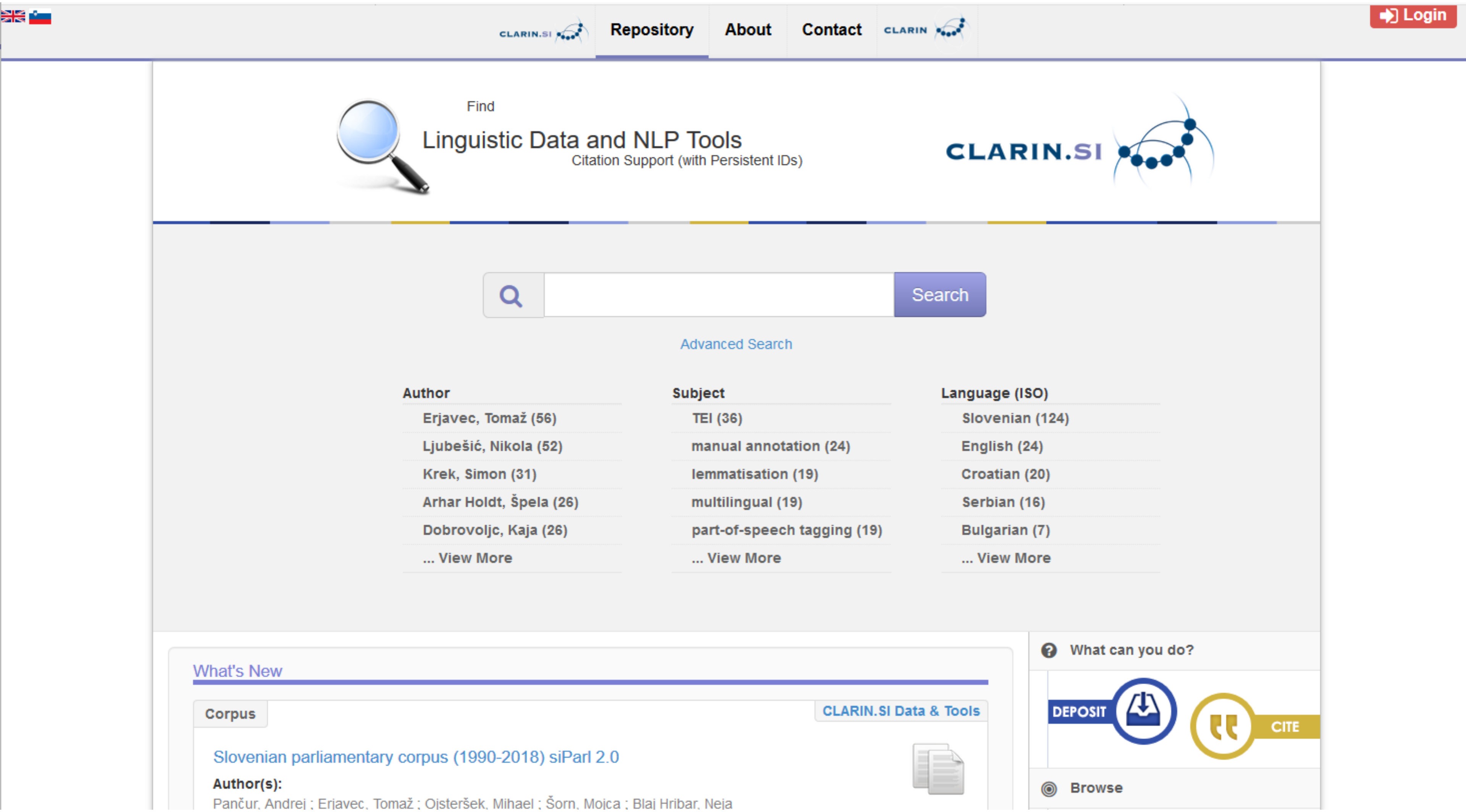
- Each repository entry is assigned a PID
- So, even if the repository platform is changed, the PIDs stay the same (but need to be reconfigured)
- Best known solution: DOI
- CLARIN used the Handle system
- **http://hdl.handle.net/11356/1044** → https://www.clarin.si/repository/xmlui/handle/11356/1044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

CLARIN.SI

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

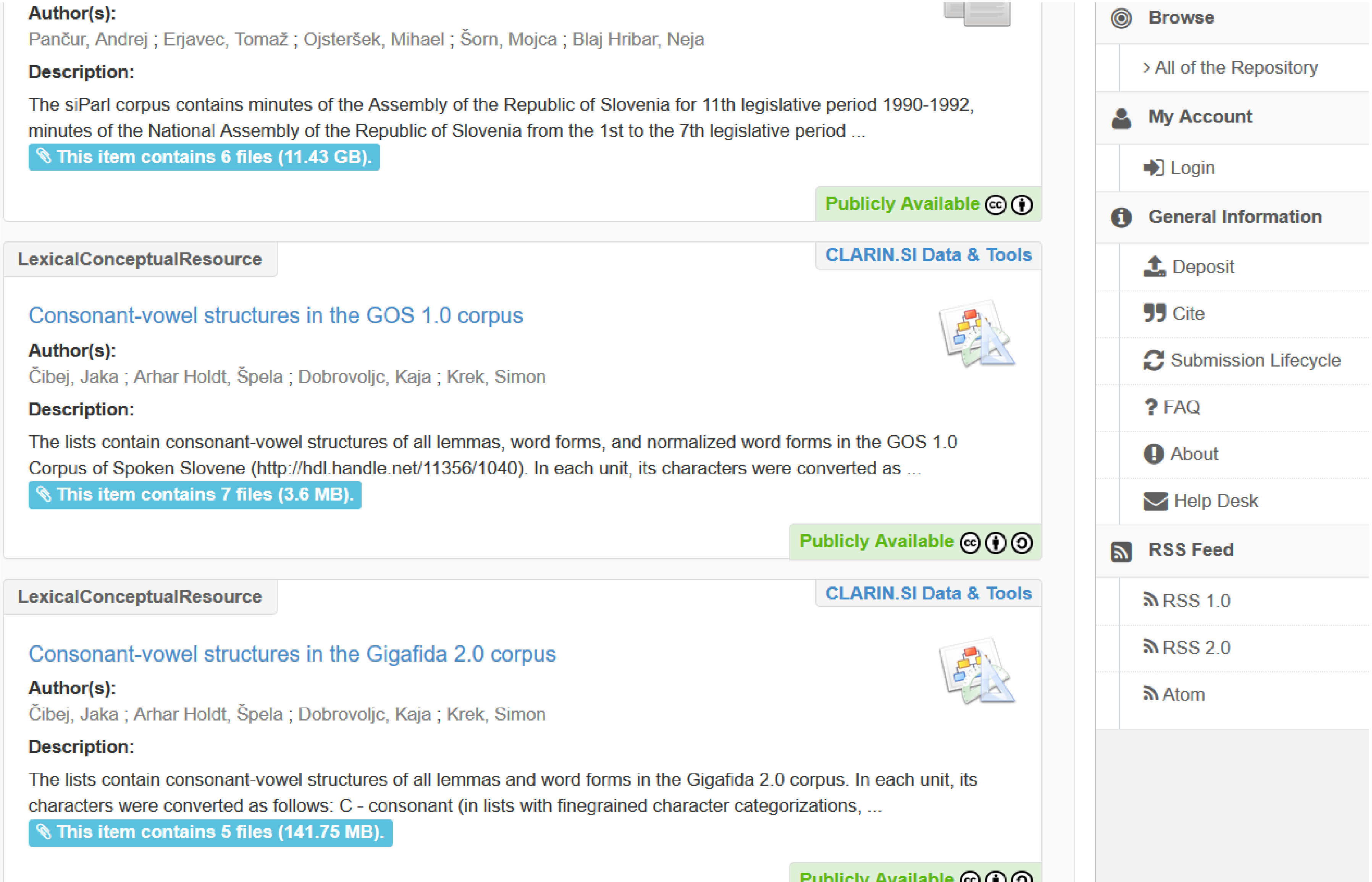
Author(s):

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:

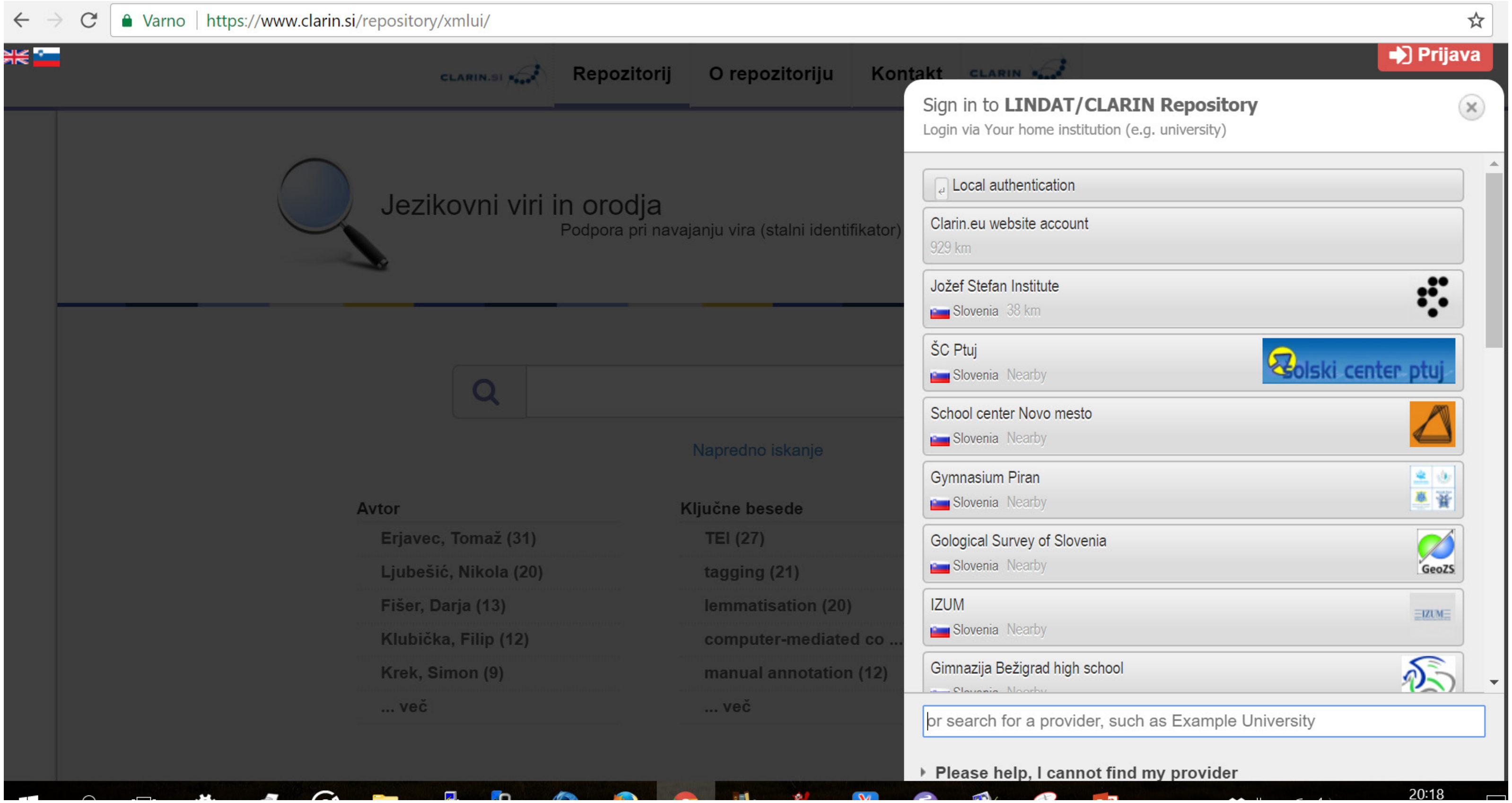
The lists contain consonant-vowel structures of all lemmas and word forms in the Gigafida 2.0 corpus. In each unit, its characters were converted as follows: C - consonant (in lists with finegrained character categorizations, ...

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



Jezikovni viri in orodja
Podpora pri navajanju vira (stalni identifikator)

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)

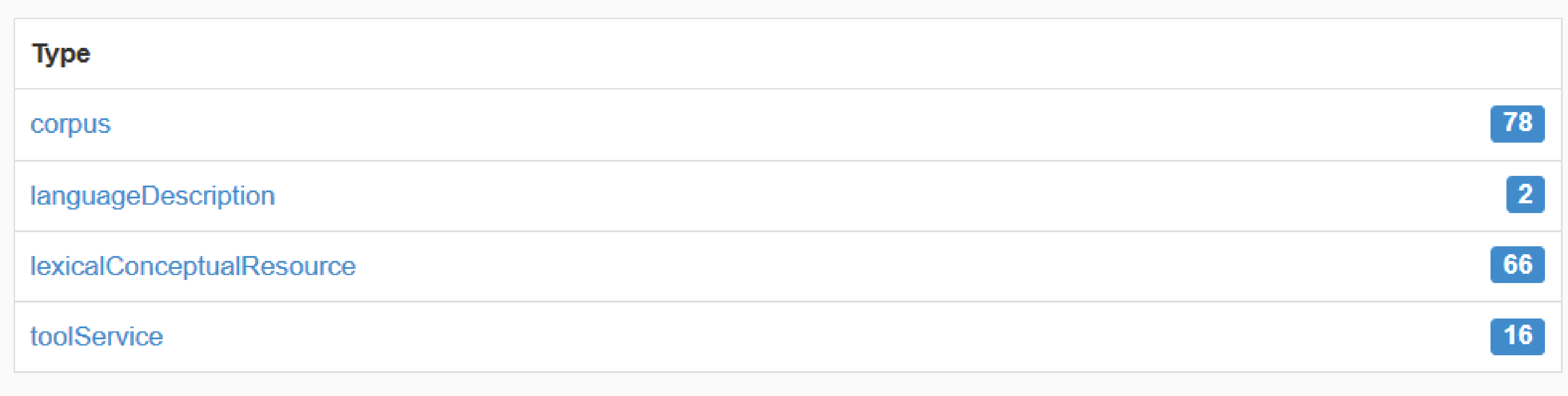
- Local authentication
- Clarin.eu website account — 929 km
- Jožef Stefan Institute — Slovenia 38 km
- ŠC Ptuj — Slovenia Nearby
- School center Novo mesto — Slovenia Nearby
- Gymnasium Piran — Slovenia Nearby
- Gological Survey of Slovenia — Slovenia Nearby
- IZUM — Slovenia Nearby
- Gimnazija Bežigrad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

or search for a provider, such as Example University

Please help, I cannot find my provider

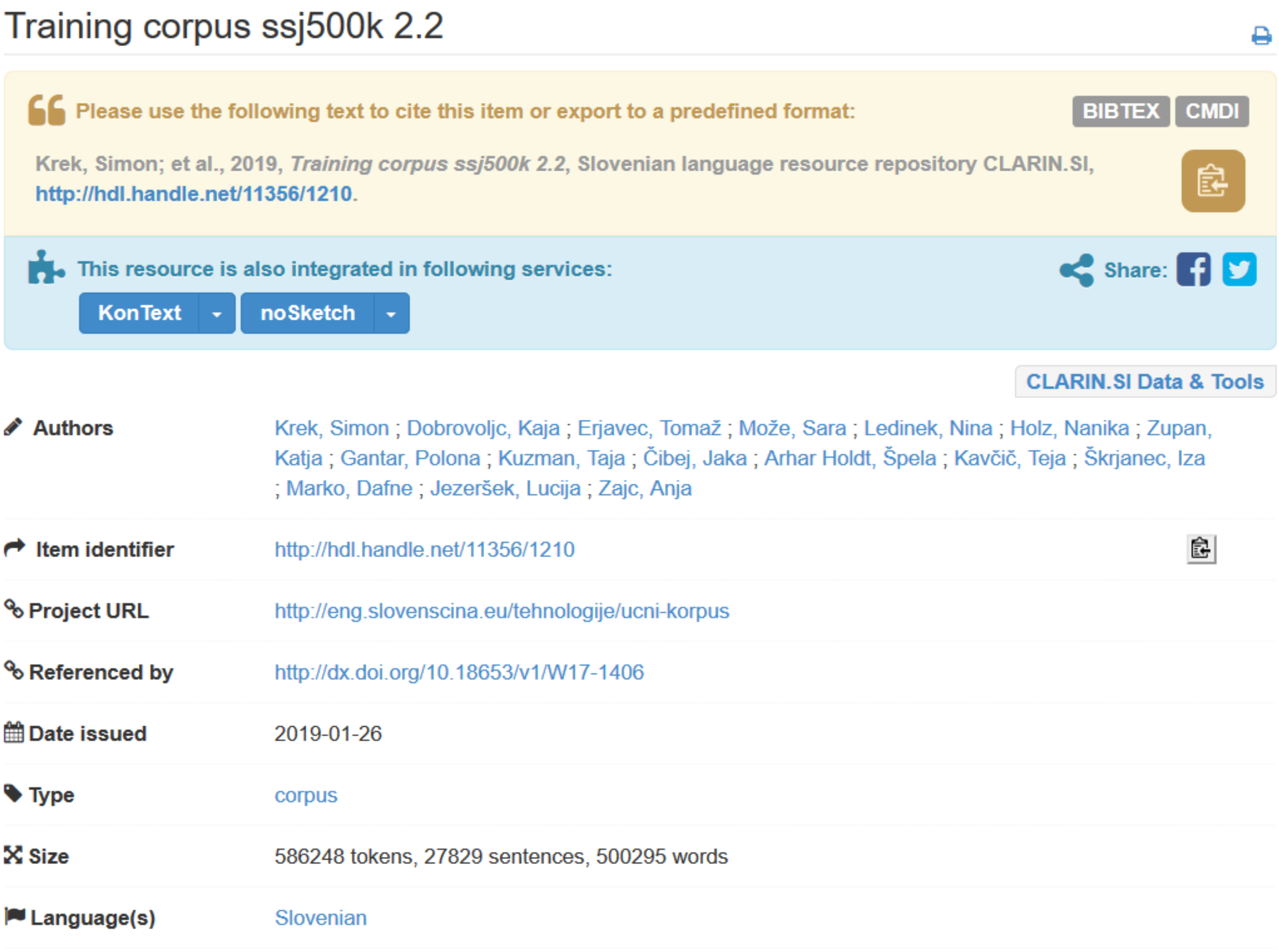
How to find interesting resources

- 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

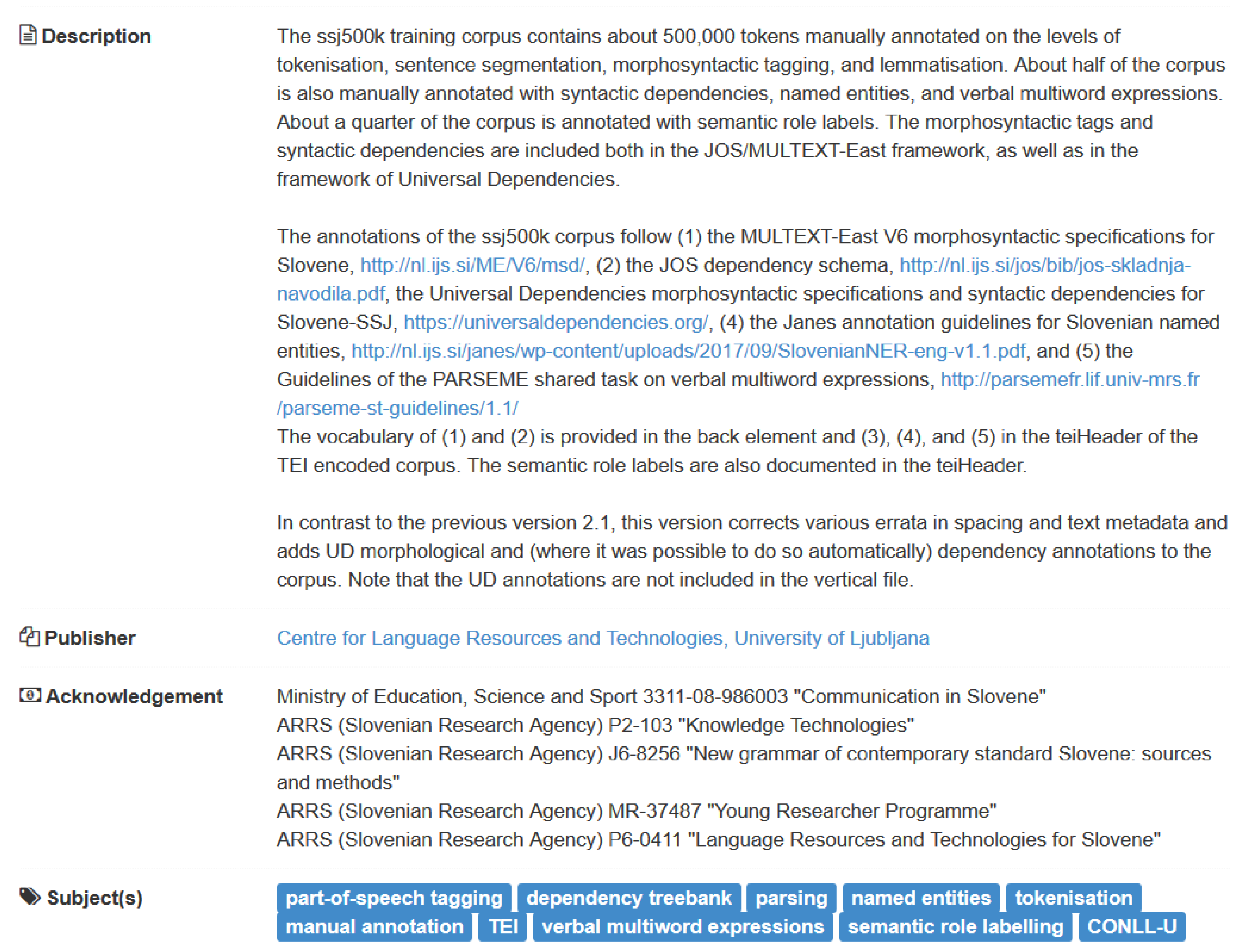
Please use the following text to cite this item or export to a predefined format:

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

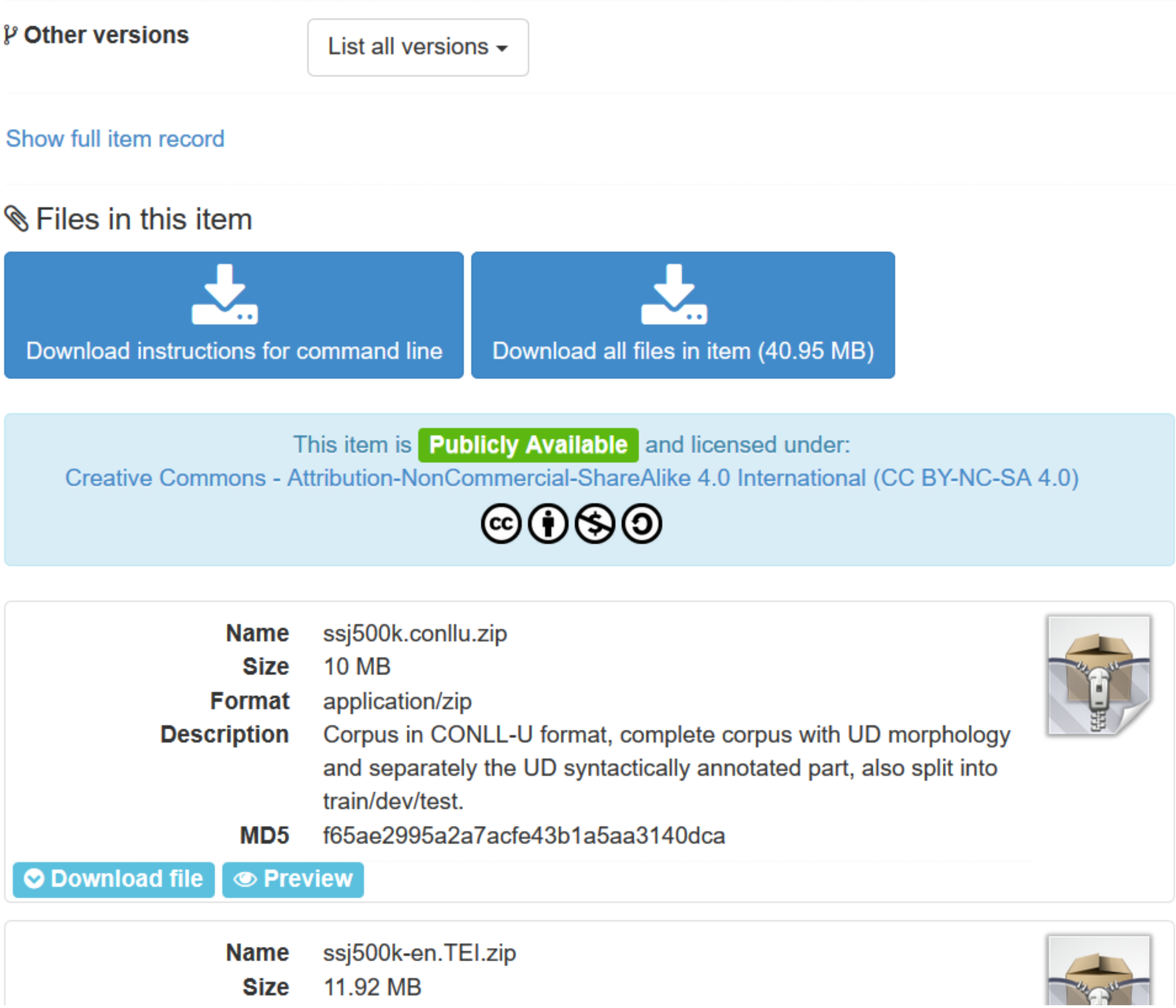
Acknowledgement

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



Other versions List all versions

Show full item record

Files in this item

Download instructions for command line

Download all files in item (40.95 MB)

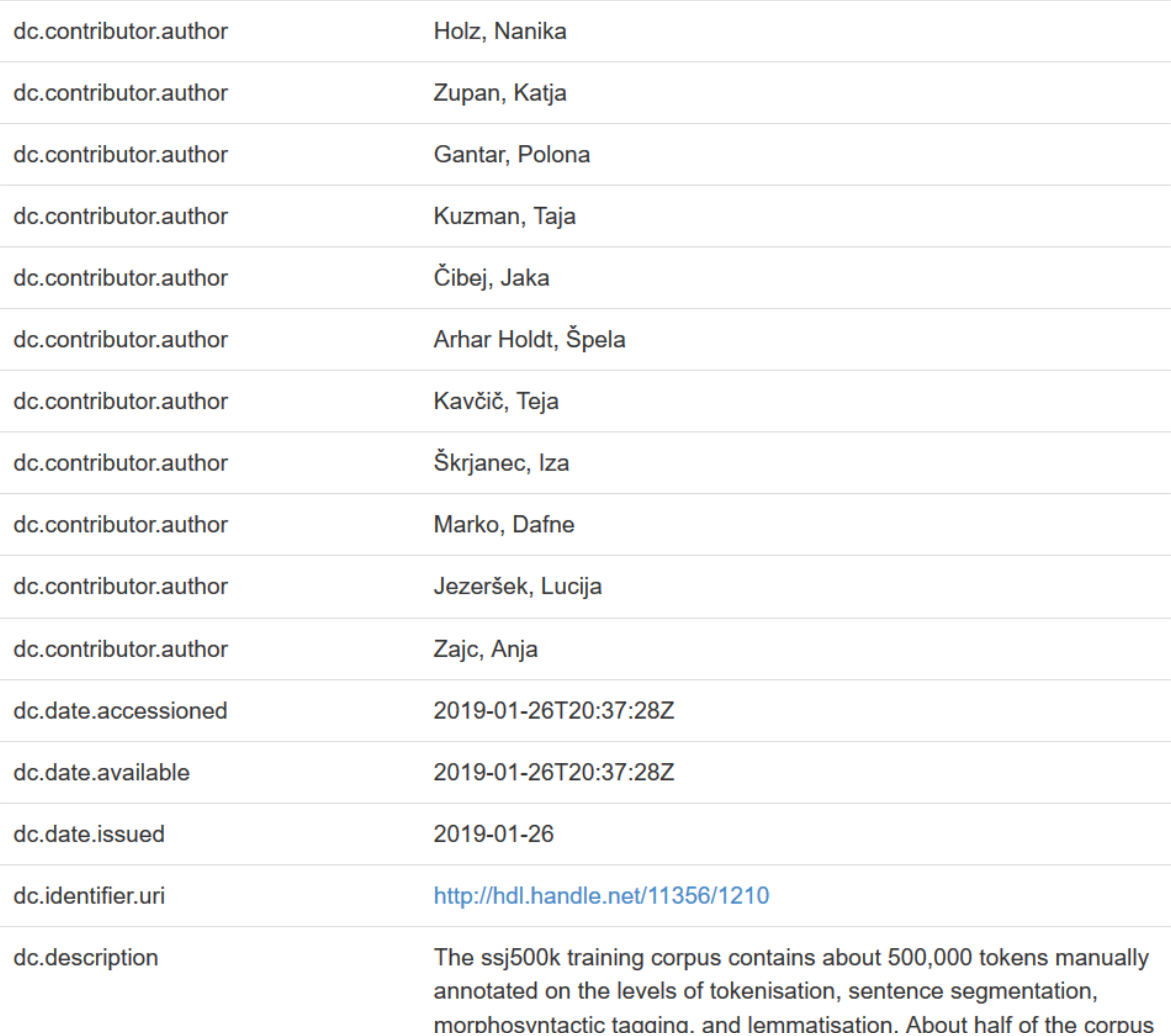
This item is **Publicly Available** and licensed under:
Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

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

Download file Preview

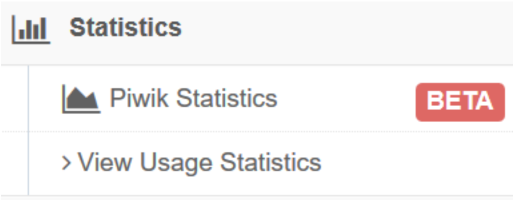
Name	ssj500k-en.TEI.zip
Size	11.92 MB

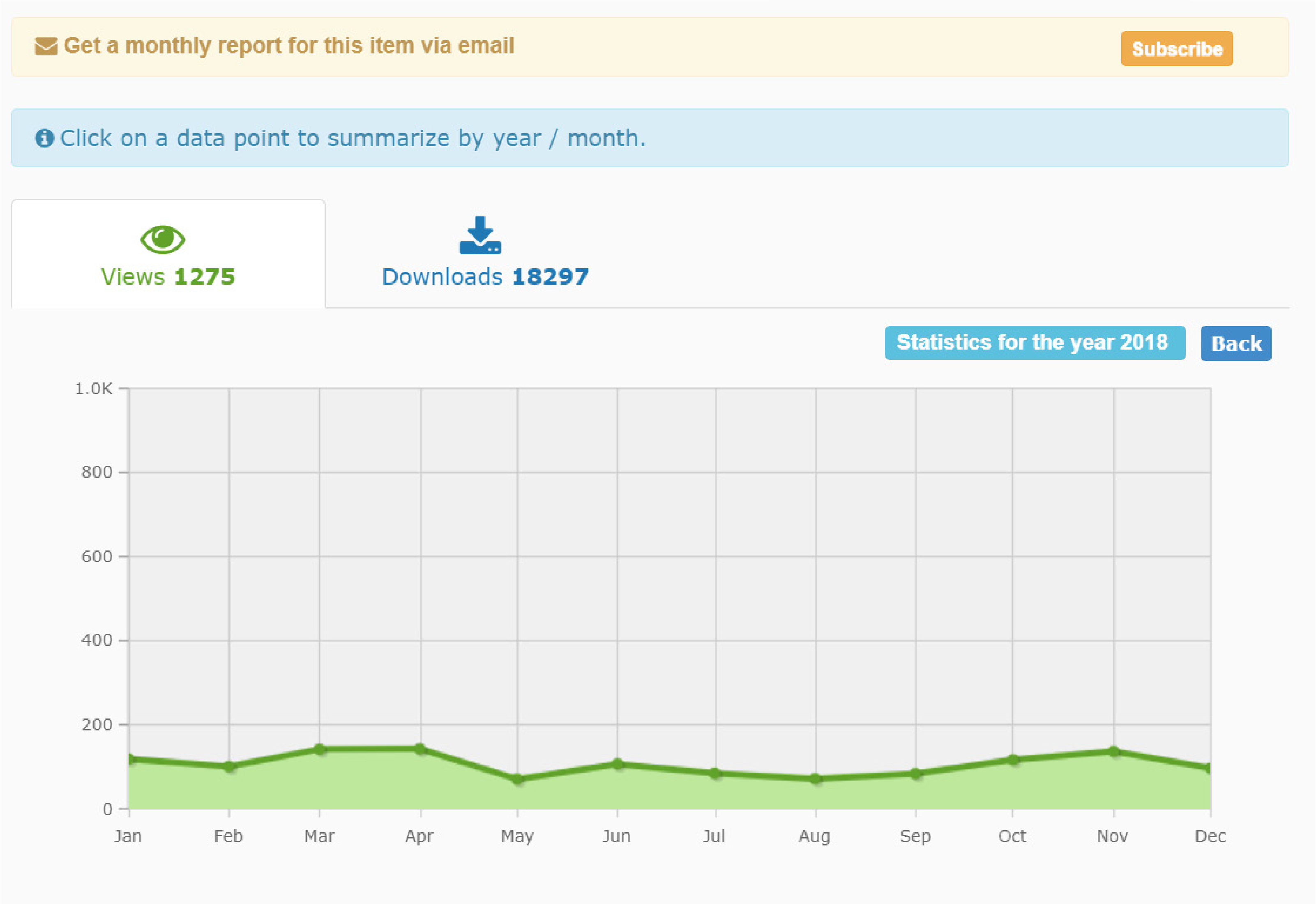
DC metadata



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

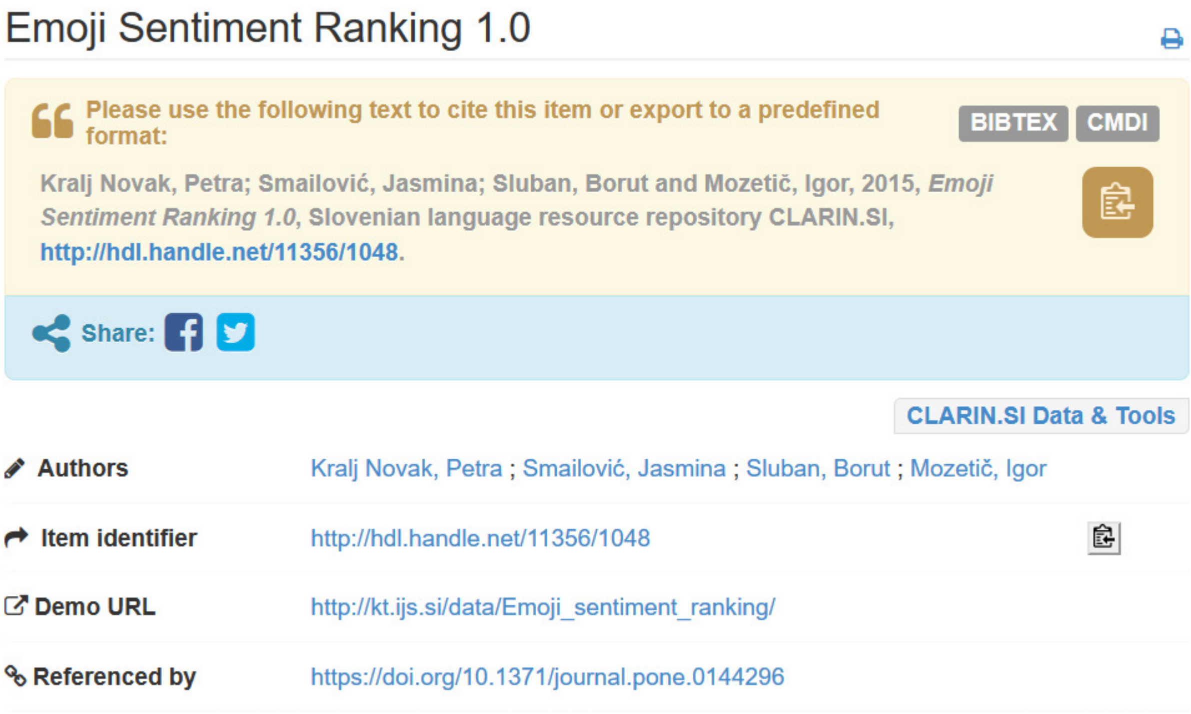



Get a monthly report for this item via email

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

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

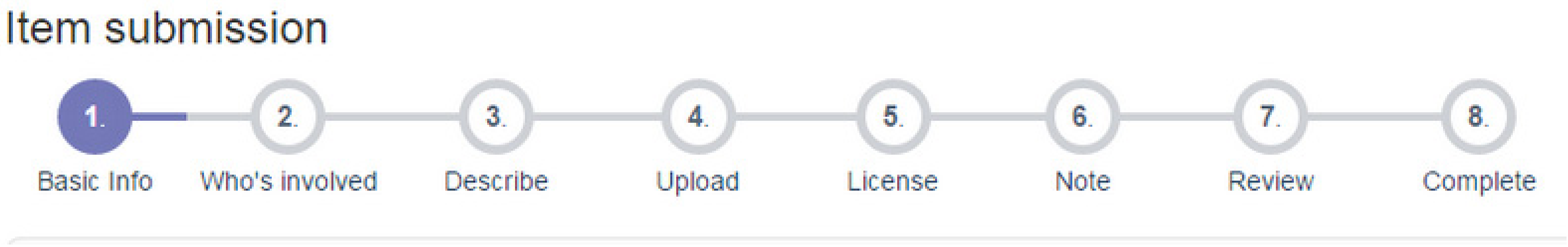
Please use the following text to cite this item or export to a predefined format:

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

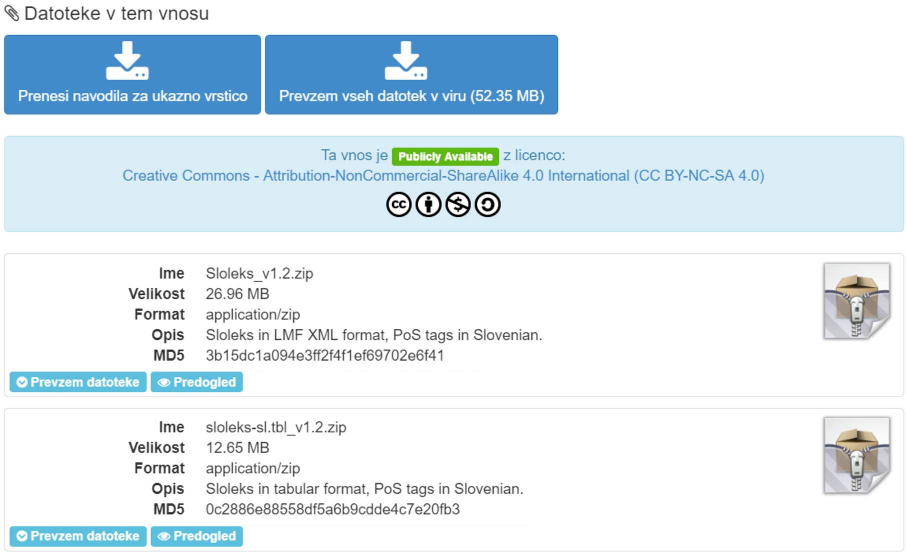
- AAI log-in
- Fairly simple workflow:


Item submission

- 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



Prenesi navodila za ukazno vrstico
Prevzem vseh datotek v viru (52.35 MB)
Publicly Available
Prevzem datoteke
Predogled
Prevzem datoteke
Predogled

Datoteke v tem vnosu

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

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?

- 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

www.slovenščina.eu

MTE

АБУМАТРАН
ABUMATRAN

IMP

ReLDI

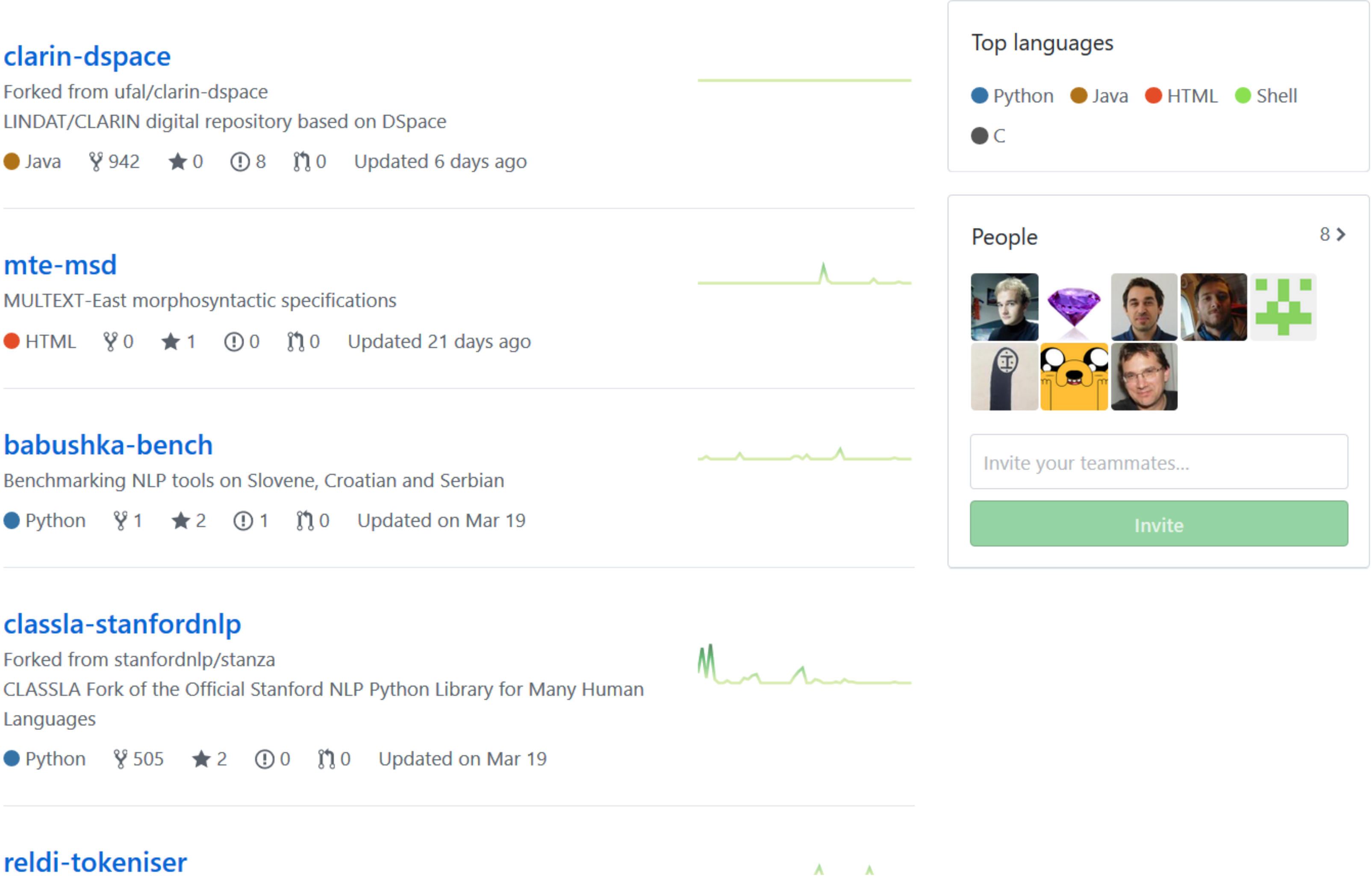
JANES

ΚΑΣ 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 >

Standardising annotation of language data: TEI

Tomaž Erjavec

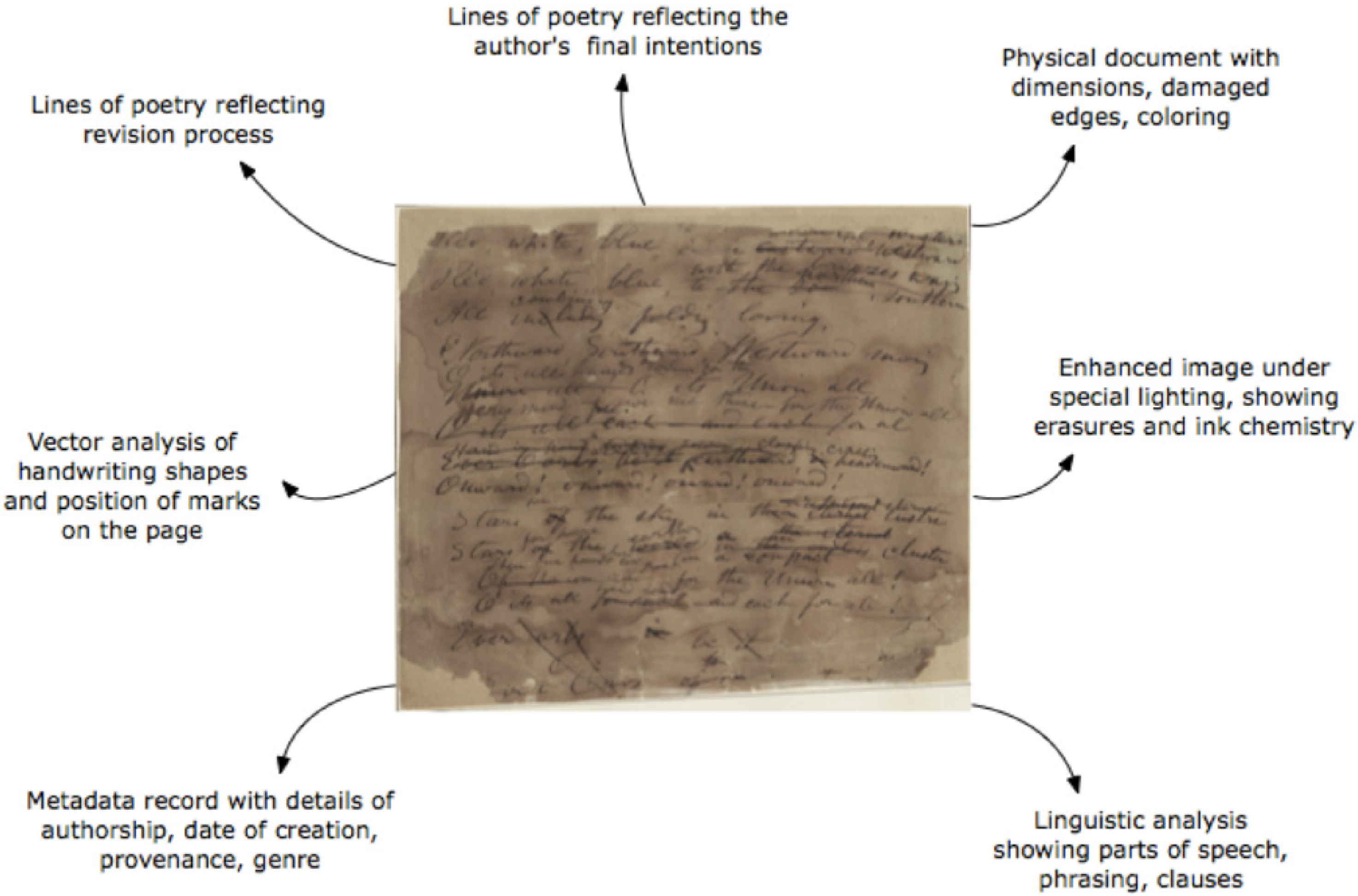
Department of Knowledge Technologies, Jožef Stefan Institute

FRI, 29th April, 2020

Overview of the lecture

- Introduction
- History, organisation and scope of TEI
- TEI modules
- TEI in Slovenia

Text Encoding: Representing Research Objects



Lines of poetry reflecting the author's final intentions

Lines of poetry reflecting revision process

Physical document with dimensions, damaged edges, coloring

Enhanced image under special lighting, showing erasures and ink chemistry

Vector analysis of handwriting shapes and position of marks on the page

Metadata record with details of authorship, date of creation, provenance, genre

Linguistic analysis showing parts of speech, phrasing, clauses

Overview of Text Encoding and the TEI, slide 1 of 6

https://www.wwp.northeastern.edu/outreach/seminars/context_2011-09/presentations/overview/overview_new_01.xhtml

Digital formats for text

- Graphic formats (facsimile of text)
- Sound formats (speech)
- Video formats (movie of a speaker)
- Tabular formats (lexicons, some training sets)
- **XML** (annotated text and pointers)

Advantages of XML

- Supports hierarchical structures and pointers
- Allows mixing text and annotations
- A W3C standard
- Formal validation via a schema:
  DTD, W3C schema, RelaxNG, Schematron
- Many associated standards:
  XInclude, XPath, XSLT, Xquery
- Tool support:
  Saxon (XSLT), eXist (XQuery), editors (Oxygen, emacs)
- “Simple” conversion to applications formats:
  XML, HTML, JSON, tables, PDF, …
- A good archive format

XML schemas for text annotation

- XML is a meta-language:
  it is the schema defines the allowed elements and attributes, and allowed nestings
- In addition to the formal schema (syntax) we also need documentation of the meaning of the elements, attributes, values (semantics)
- We can develop a schema for a particular type of documents ourselves
- For data interchange (and tool support) it is better to use a standard schema
- For linguistic annotation there are (too) many standard and best-practice schemas:
  - ISO TC 37 SC4 standards for encoding language resources
  - National projects (TCF: Germany, FoLiA: Netherlands, …)
  - and TEI, which is, however, much broader

TEI < Text Encoding Initiative >

- Started as a research project in the humanities
  - supported by three professional associations
  - financed 1990-1994 (ZDA, EU)
- influential text types:
  - digital libraries, text collections
  - language corpora
  - scholarly datasets
- International consortium established 1999
- Web page: http://www.tei-c.org/

Goals of the TEI

- Better interchange and integration of scholarly data
- Support for all texts, in all languages, from all periods
- Guidance for the perplexed: what to encode:
  a user-driven codification of existing best practice
- Assistance for the specialist: how to encode:
  a loose framework into which unpredictable extensions can be fitted

These apparently incompatible goals result in a highly flexible, modular, environment

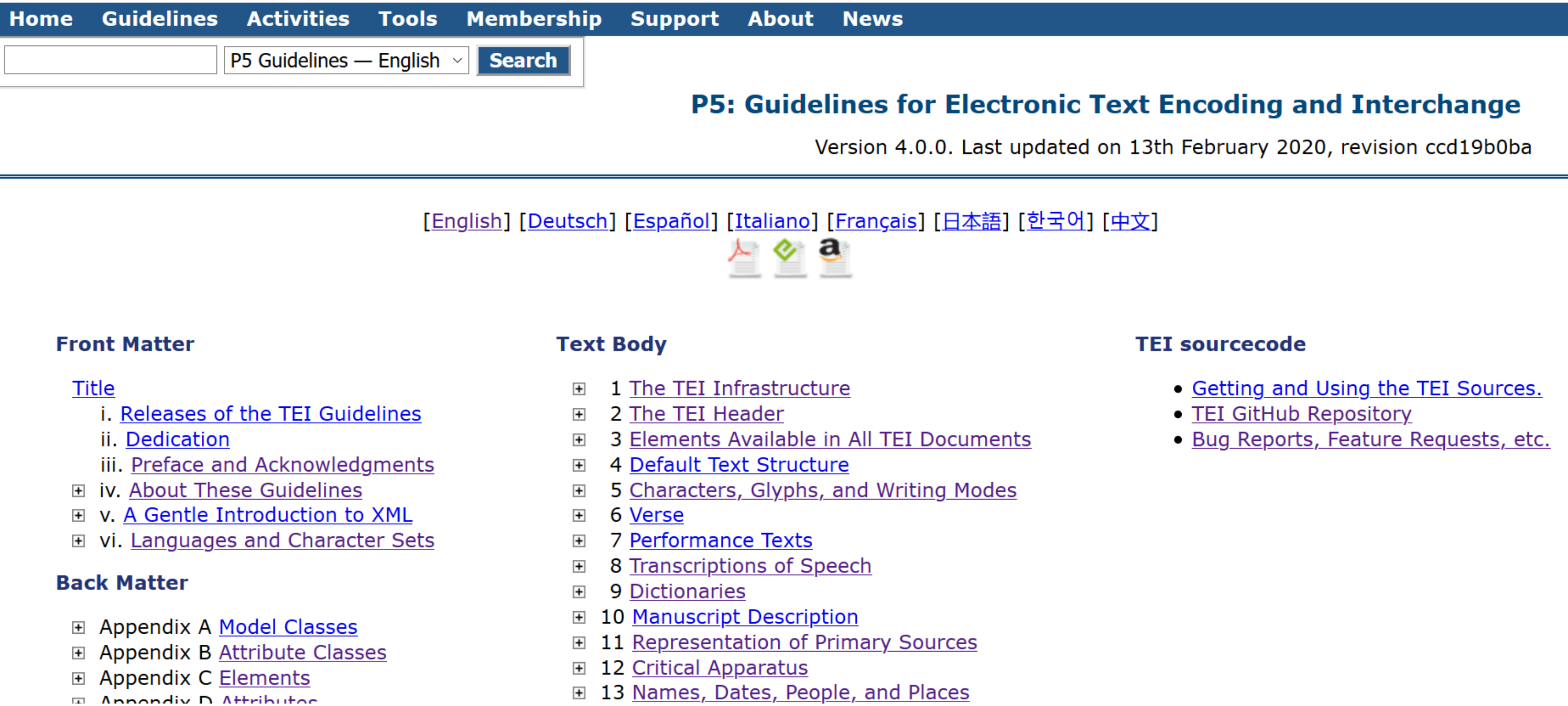
Legacy of the TEI

- A way of looking at what ‘text’ really is
- A codification of current scholarly practice
- A set of shared assumptions and priorities about the digital agenda:
    - focus on content and function (rather than presentation)
    - identify generic solutions (rather than application-specific ones)

TEI Guidelines

- “TEI Guidelines for Electronic Text Encoding and Interchange” cf. http://www.tei-c.org/Guidelines/
- Cover generic structures as well as very specific fields
- A large collection of XML element and attribute definitions
- TEI Guidelines comprise documentation and formal definitions
Written in TEI ODD: One Document Does it all
- In print ~1.200 pages
- A modular system for creating personalized schemas
- Maintained on GitHub,
available in TEI/XML, HTML, PDF, EPUB,…

TEI Guidelines on the Web



P5: Guidelines for Electronic Text Encoding and Interchange

Version 4.0.0. Last updated on 13th February 2020, revision ccd19b0ba

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [한국어] [中文]

Front Matter

Title
- i. Releases of the TEI Guidelines
- ii. Dedication
- iii. Preface and Acknowledgments
- iv. About These Guidelines
- v. A Gentle Introduction to XML
- vi. Languages and Character Sets

Back Matter

- Appendix A Model Classes
- Appendix B Attribute Classes
- Appendix C Elements
- Appendix D Attributes

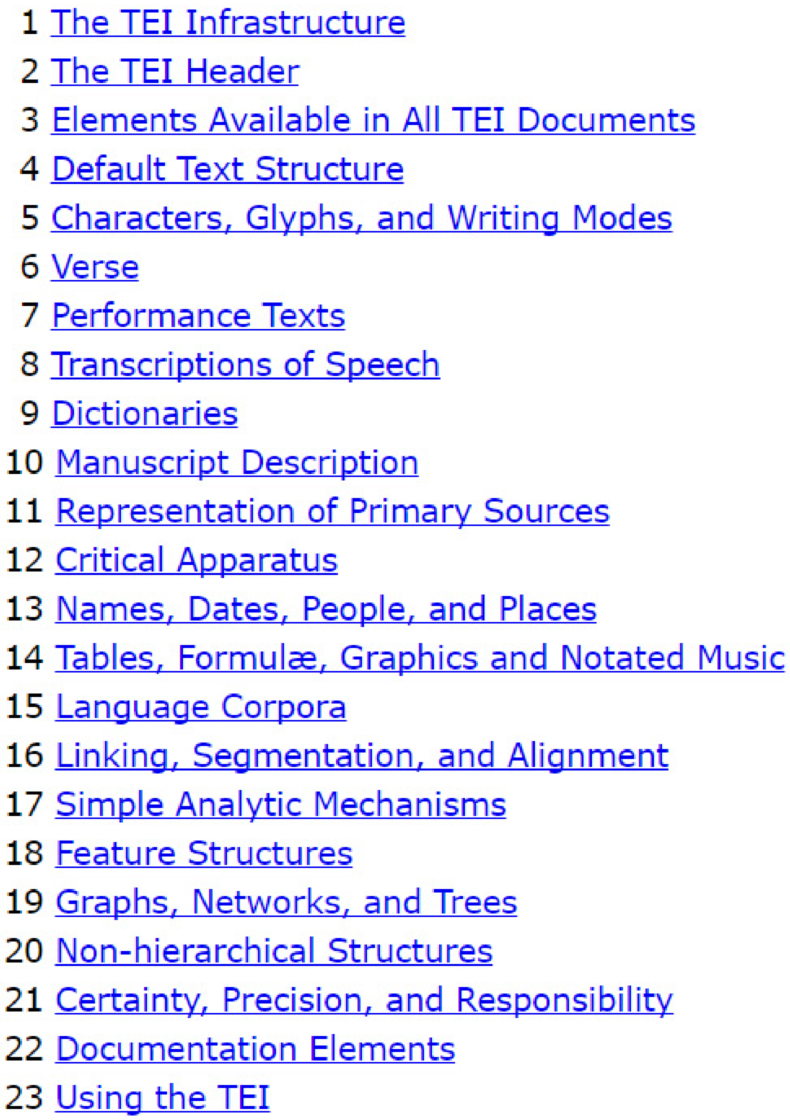
Text Body

- 1 The TEI Infrastructure
- 2 The TEI Header
- 3 Elements Available in All TEI Documents
- 4 Default Text Structure
- 5 Characters, Glyphs, and Writing Modes
- 6 Verse
- 7 Performance Texts
- 8 Transcriptions of Speech
- 9 Dictionaries
- 10 Manuscript Description
- 11 Representation of Primary Sources
- 12 Critical Apparatus
- 13 Names, Dates, People, and Places

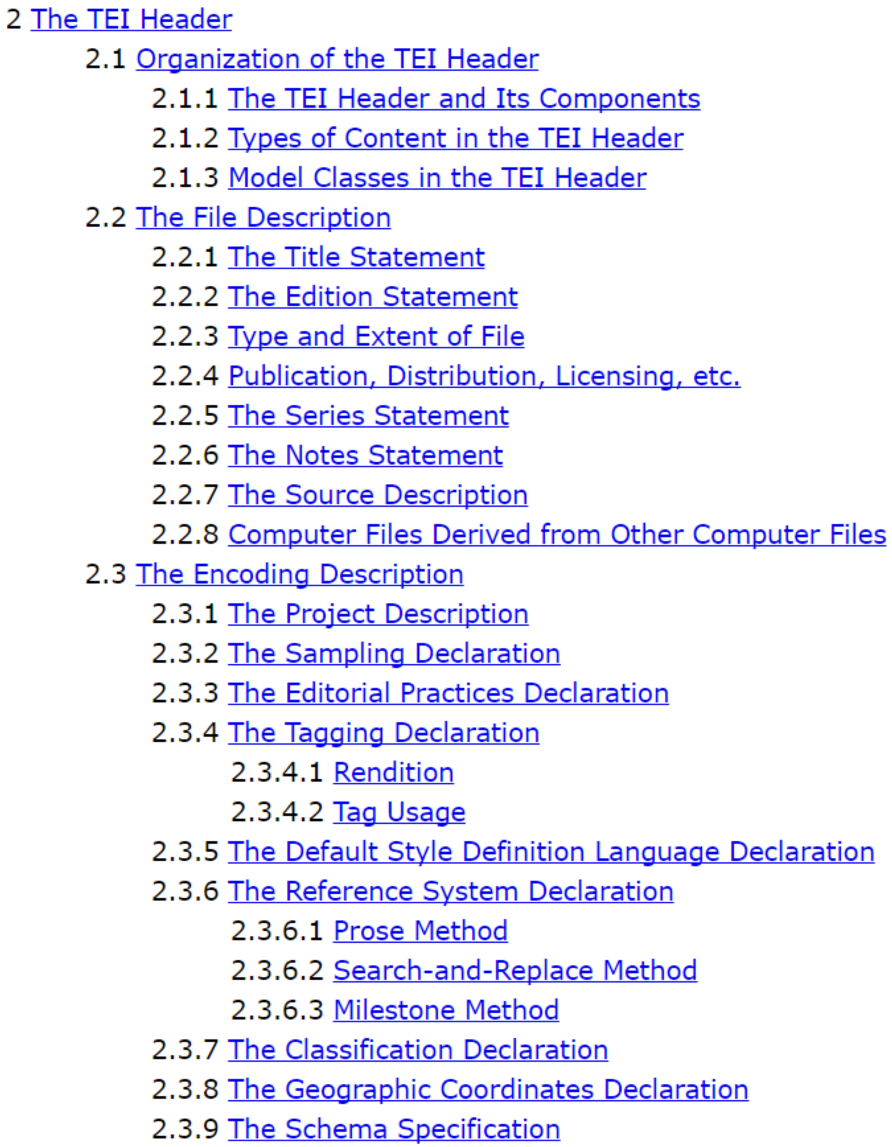
TEI sourcecode

- Getting and Using the TEI Sources.
- TEI GitHub Repository
- Bug Reports, Feature Requests, etc.

Chapters / modules



1 The TEI Infrastructure
2 The TEI Header
3 Elements Available in All TEI Documents
4 Default Text Structure
5 Characters, Glyphs, and Writing Modes
6 Verse
7 Performance Texts
8 Transcriptions of Speech
9 Dictionaries
10 Manuscript Description
11 Representation of Primary Sources
12 Critical Apparatus
13 Names, Dates, People, and Places
14 Tables, Formulæ, Graphics and Notated Music
15 Language Corpora
16 Linking, Segmentation, and Alignment
17 Simple Analytic Mechanisms
18 Feature Structures
19 Graphs, Networks, and Trees
20 Non-hierarchical Structures
21 Certainty, Precision, and Responsibility
22 Documentation Elements
23 Using the TEI



2 The TEI Header
- 2.1 Organization of the TEI Header
  - 2.1.1 The TEI Header and Its Components
  - 2.1.2 Types of Content in the TEI Header
  - 2.1.3 Model Classes in the TEI Header
- 2.2 The File Description
  - 2.2.1 The Title Statement
  - 2.2.2 The Edition Statement
  - 2.2.3 Type and Extent of File
  - 2.2.4 Publication, Distribution, Licensing, etc.
  - 2.2.5 The Series Statement
  - 2.2.6 The Notes Statement
  - 2.2.7 The Source Description
  - 2.2.8 Computer Files Derived from Other Computer Files
- 2.3 The Encoding Description
  - 2.3.1 The Project Description
  - 2.3.2 The Sampling Declaration
  - 2.3.3 The Editorial Practices Declaration
  - 2.3.4 The Tagging Declaration
    - 2.3.4.1 Rendition
    - 2.3.4.2 Tag Usage
  - 2.3.5 The Default Style Definition Language Declaration
  - 2.3.6 The Reference System Declaration
    - 2.3.6.1 Prose Method
    - 2.3.6.2 Search-and-Replace Method
    - 2.3.6.3 Milestone Method
  - 2.3.7 The Classification Declaration
  - 2.3.8 The Geographic Coordinates Declaration
  - 2.3.9 The Schema Specification

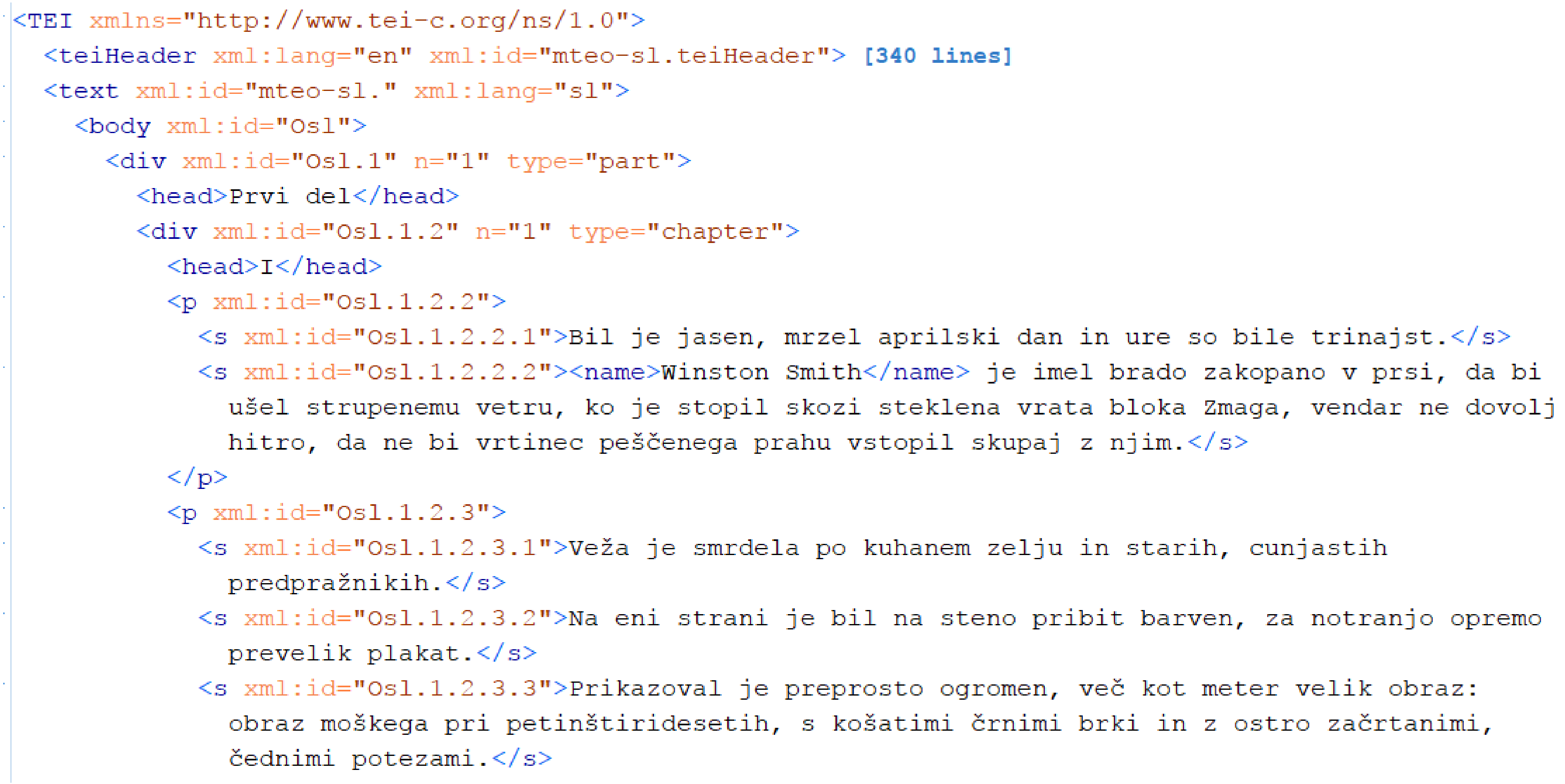
The TEI ecosystem

- TEI Consortium (Technical Council, Board of Directors)
- Open source and collaborative development
- Guidelines and web pages regularly maintained
- A friendly and active tei-l mailing list
- jTEI journal and annual conferences of the TEI
- Support tools:
  - Roma: Web interface to parametrise TEI and generate XML schema
  - TEI Stylesheets: conversion to and from TEI for many formats (docx, open office, html, markdown, ePub, …)

TEI in Slovenia (1998-)

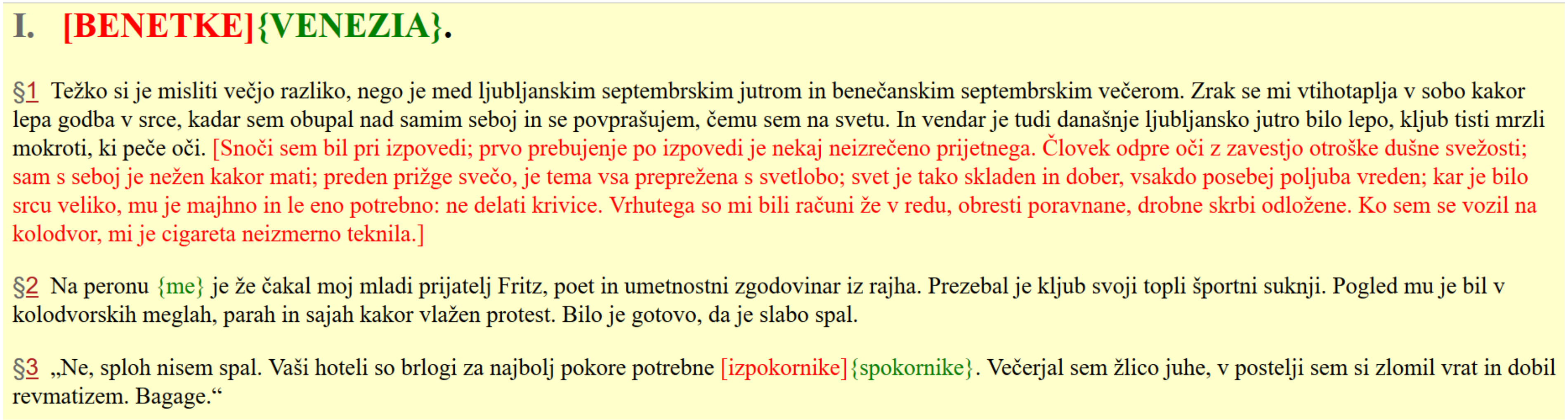
- Reference corpora:
  Fida, FidaPLUS, Gigafida; ssj500k; hr500k, …
- EU projects (IJS et al.): MULTEXT-East, IMPACT, ELEXIS
- National projects ZRC SAZU + IJS:
  Scholarly editions of Slovenian literature, Slovenian biography
- Projects with UL Dept. for Asian languages + IJS:
  Japanese-Slovene learners' dictionary and corpora
- DARIAH-SI + CLARIN.SI:
  digital library + corpora (e.g. siParl)
- National projects IJS:
  - JOS: Linguistic Annotation of Slovene
  - Janes: Linguistic Annotation of Non-Standard Slovene
  - KAS: Slovene Scientific Texts: Resources and Description

Examples of TEI use: “1984”



```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader xml:lang="en" xml:id="mteo-sl.teiHeader"> [340 lines]
  <text xml:id="mteo-sl." xml:lang="sl">
    <body xml:id="Osl">
      <div xml:id="Osl.1" n="1" type="part">
        <head>Prvi del</head>
        <div xml:id="Osl.1.2" n="1" type="chapter">
          <head>I</head>
          <p xml:id="Osl.1.2.2">
            <s xml:id="Osl.1.2.2.1">Bil je jasen, mrzel aprilski dan in ure so bile trinajst.</s>
            <s xml:id="Osl.1.2.2.2"><name>Winston Smith</name> je imel brado zakopano v prsi, da bi
              ušel strupenemu vetru, ko je stopil skozi steklena vrata bloka Zmaga, vendar ne dovolj
              hitro, da ne bi vrtinec peščenega prahu vstopil skupaj z njim.</s>
          </p>
          <p xml:id="Osl.1.2.3">
            <s xml:id="Osl.1.2.3.1">Veža je smrdela po kuhanem zelju in starih, cunjastih
              predpražnikih.</s>
            <s xml:id="Osl.1.2.3.2">Na eni strani je bil na steno pribit barven, za notranjo opremo
              prevelik plakat.</s>
            <s xml:id="Osl.1.2.3.3">Prikazoval je preprosto ogromen, več kot meter velik obraz:
              obraz moškega pri petinštiridesetih, s košatimi črnimi brki in z ostro začrtanimi,
              čednimi potezami.</s>
```

Izidor Cankar: S poti



I. [BENETKE]{VENEZIA}.

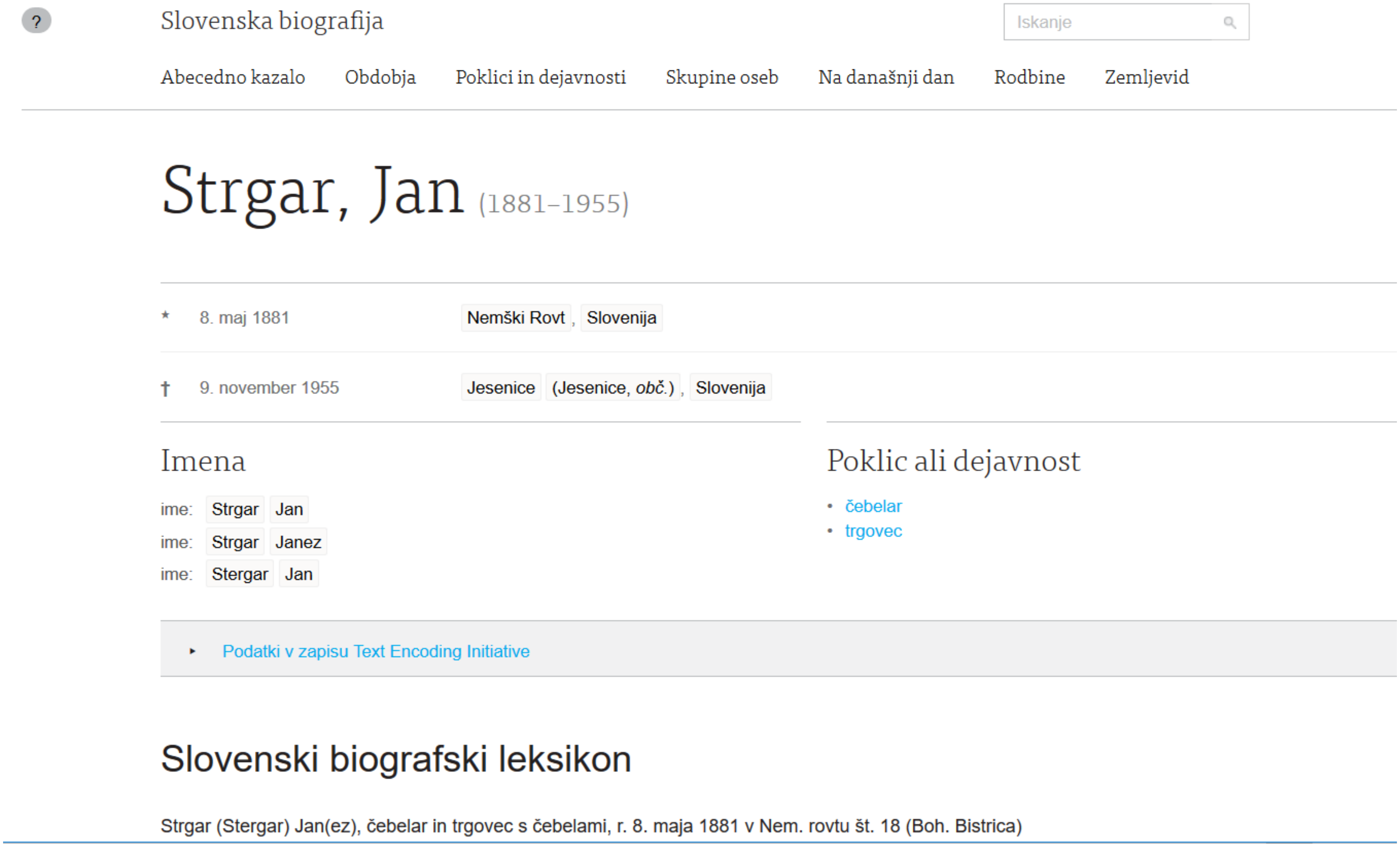
§1 Težko si je misliti večjo razliko, nego je med ljubljanskim septembrskim jutrom in benečanskim septembrskim večerom. Zrak se mi vtihotaplja v sobo kakor lepa godba v srce, kadar sem obupal nad samim seboj in se povprašujem, čemu sem na svetu. In vendar je tudi današnje ljubljansko jutro bilo lepo, kljub tisti mrzli mokroti, ki peče oči. [Snoči sem bil pri izpovedi; prvo prebujenje po izpovedi je nekaj neizrečeno prijetnega. Človek odpre oči z zavestjo otroške dušne svežosti; sam s seboj je nežen kakor mati; preden prižge svečo, je tema vsa preprežena s svetlobo; svet je tako skladen in dober, vsakdo posebej poljuba vreden; kar je bilo srcu veliko, mu je majhno in le eno potrebno: ne delati krivice. Vrhutega so mi bili računi že v redu, obresti poravnane, drobne skrbi odložene. Ko sem se vozil na kolodvor, mi je cigareta neizmerno teknila.]

§2 Na peronu {me} je že čakal moj mladi prijatelj Fritz, poet in umetnostni zgodovinar iz rajha. Prezebal je kljub svoji topli športni suknji. Pogled mu je bil v kolodvorskih meglah, parah in sajah kakor vlažen protest. Bilo je gotovo, da je slabo spal.

§3 „Ne, sploh nisem spal. Vaši hoteli so brlogi za najbolj pokore potrebne [izpokornike]{spokornike}. Večerjal sem žlico juhe, v postelji sem si zlomil vrat in dobil revmatizem. Bagage.“

```
<p id="p.3" n="3">„Ne, sploh nisem spal. Vaši hoteli so brlogi za najbolj
    pokore potrebne <app>
        <rdg wit="DS">izpokornike</rdg>
        <rdg wit="KN">spokornike</rdg>
    </app>. Večerjal sem žlico juhe, v postelji sem si zlomil vrat in dobil
    revmatizem. Bagage.“</p>
```

Slovene biography: on the web



Slovenska biografija

Iskanje

Abecedno kazalo Obdobja Poklici in dejavnosti Skupine oseb Na današnji dan Rodbine Zemljevid

Strgar, Jan (1881–1955)

*	8. maj 1881	Nemški Rovt , Slovenija
†	9. november 1955	Jesenice (Jesenice, *obč.*) , Slovenija

Imena

ime: Strgar Jan
ime: Strgar Janez
ime: Stergar Jan

Poklic ali dejavnost

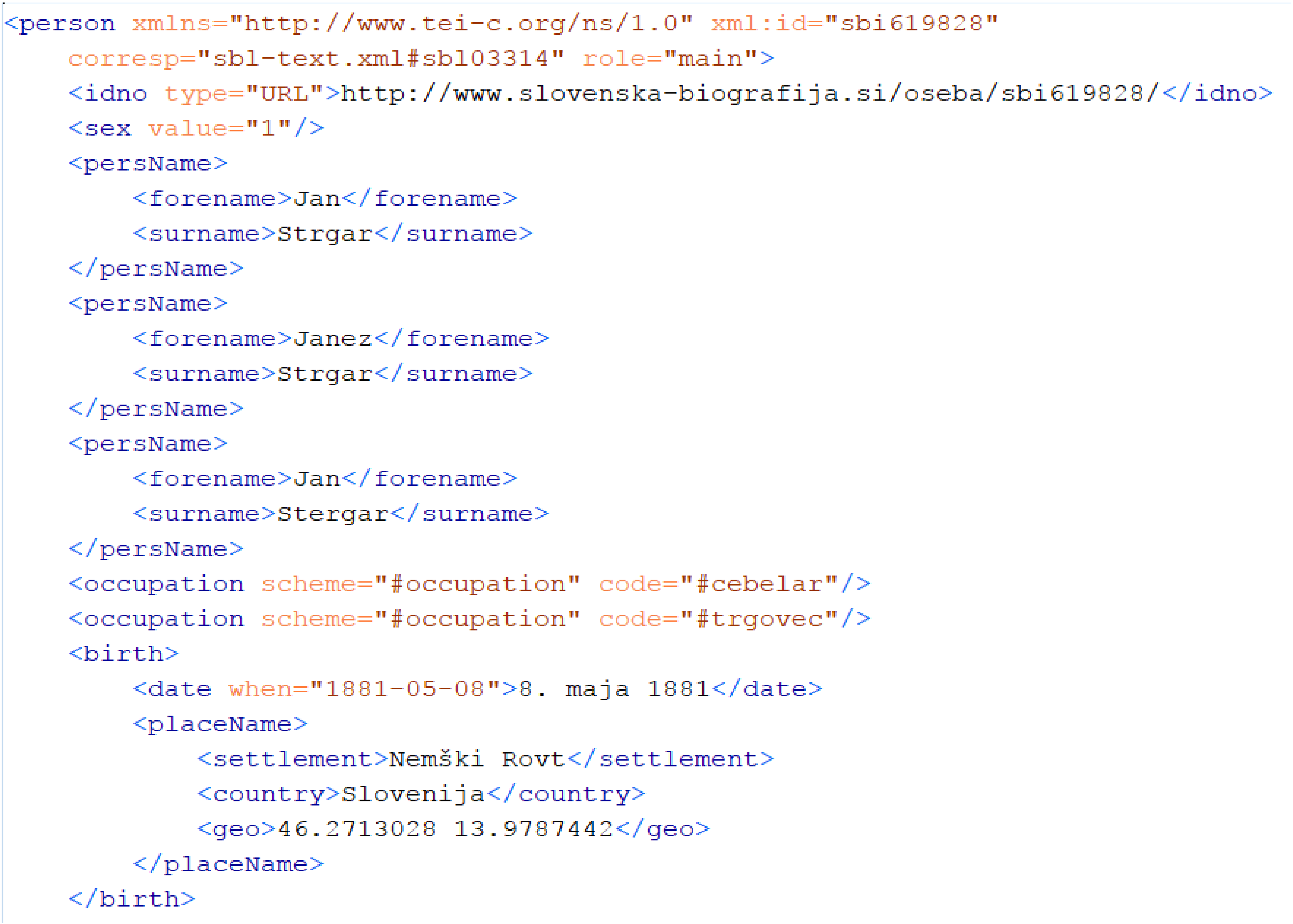
- čebelar
- trgovec

▸ Podatki v zapisu Text Encoding Initiative

Slovenski biografski leksikon

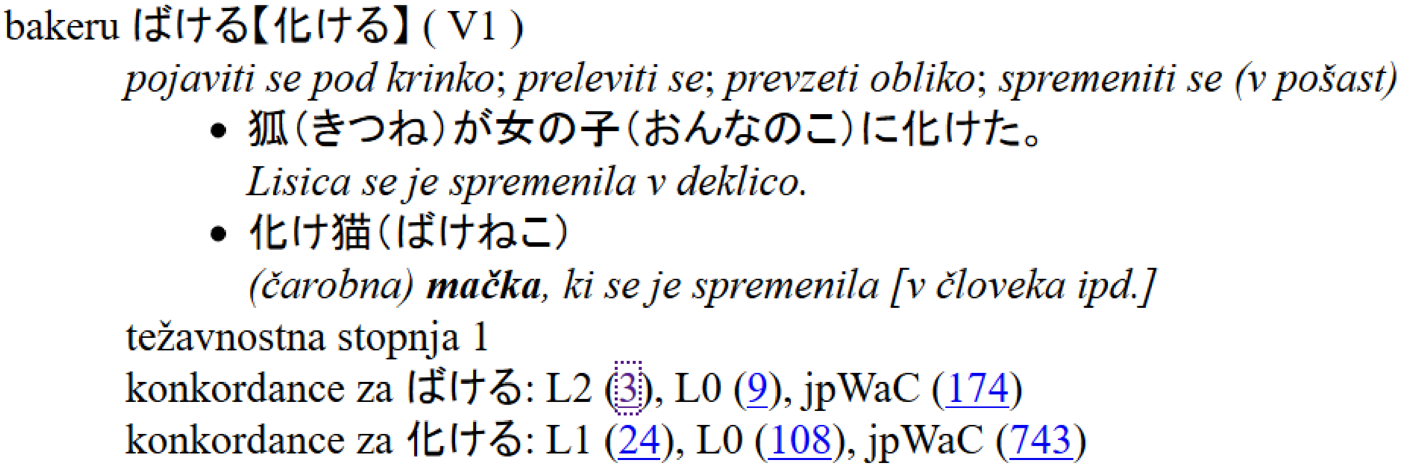
Strgar (Stergar) Jan(ez), čebelar in trgovec s čebelami, r. 8. maja 1881 v Nem. rovtu št. 18 (Boh. Bistrica)

Slovene biography: TEI



```xml
<person xmlns="http://www.tei-c.org/ns/1.0" xml:id="sbi619828"
    corresp="sbl-text.xml#sbl03314" role="main">
    <idno type="URL">http://www.slovenska-biografija.si/oseba/sbi619828/</idno>
    <sex value="1"/>
    <persName>
        <forename>Jan</forename>
        <surname>Strgar</surname>
    </persName>
    <persName>
        <forename>Janez</forename>
        <surname>Strgar</surname>
    </persName>
    <persName>
        <forename>Jan</forename>
        <surname>Stergar</surname>
    </persName>
    <occupation scheme="#occupation" code="#cebelar"/>
    <occupation scheme="#occupation" code="#trgovec"/>
    <birth>
        <date when="1881-05-08">8. maja 1881</date>
        <placeName>
            <settlement>Nemški Rovt</settlement>
            <country>Slovenija</country>
            <geo>46.2713028 13.9787442</geo>
        </placeName>
    </birth>
```

jaSlo: Japanese-Slovene dictionary



bakeru ばける【化ける】(V1)

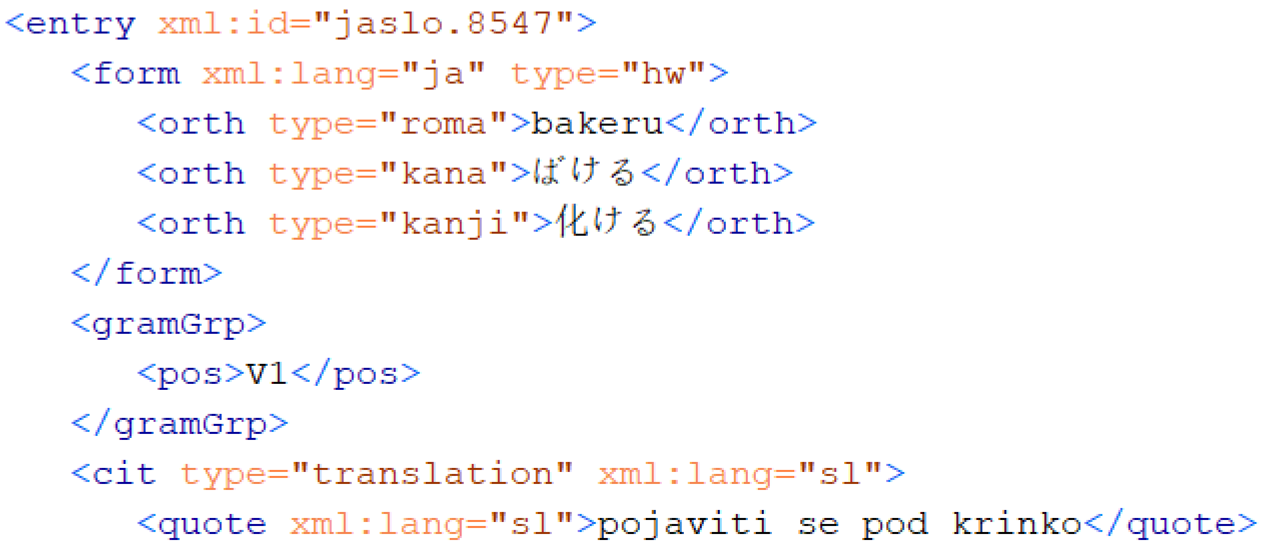
pojaviti se pod krinko; preleviti se; prevzeti obliko; spremeniti se (v pošast)

- 狐(きつね)が女の子(おんなのこ)に化けた。
  Lisica se je spremenila v deklico.
- 化け猫(ばけねこ)
  (čarobna) ***mačka****, ki se je spremenila [v človeka ipd.]*

težavnostna stopnja 1

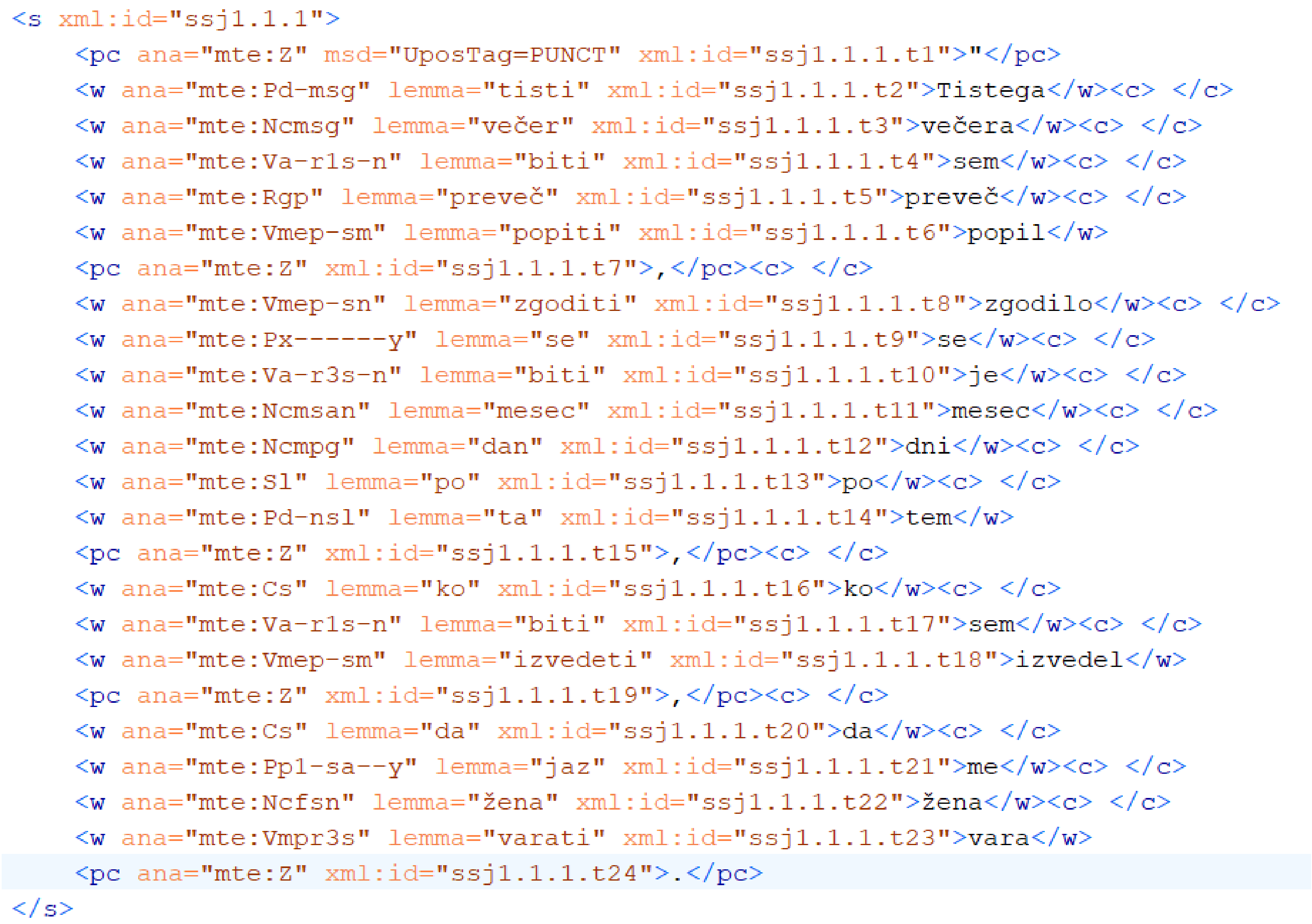
konkordance za ばける: L2 (3), L0 (9), jpWaC (174)

konkordance za 化ける: L1 (24), L0 (108), jpWaC (743)



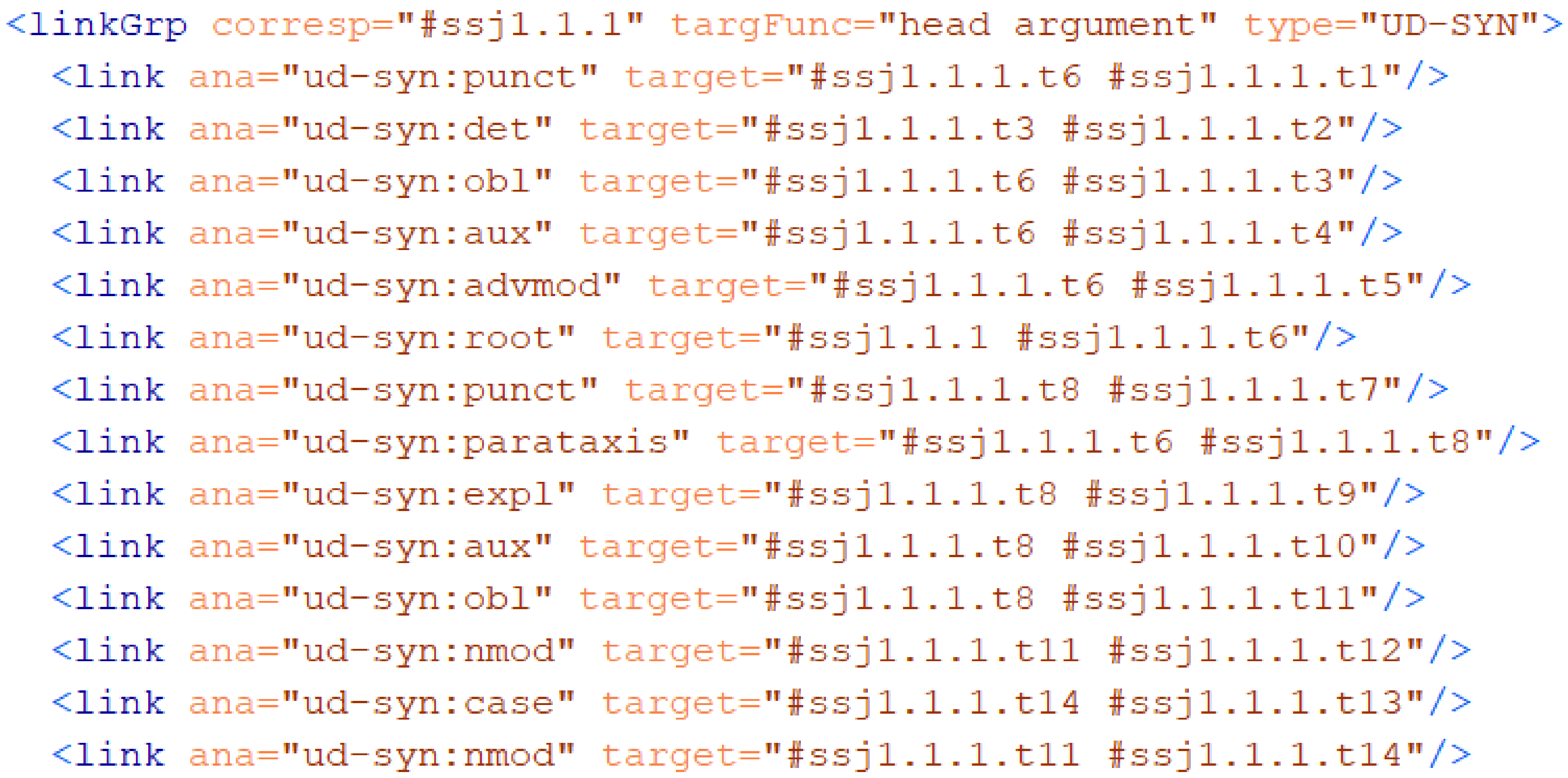
```
<entry xml:id="jaslo.8547">
   <form xml:lang="ja" type="hw">
      <orth type="roma">bakeru</orth>
      <orth type="kana">ばける</orth>
      <orth type="kanji">化ける</orth>
   </form>
   <gramGrp>
      <pos>V1</pos>
   </gramGrp>
   <cit type="translation" xml:lang="sl">
      <quote xml:lang="sl">pojaviti se pod krinko</quote>
```

ssj500k v2.2: training corpus



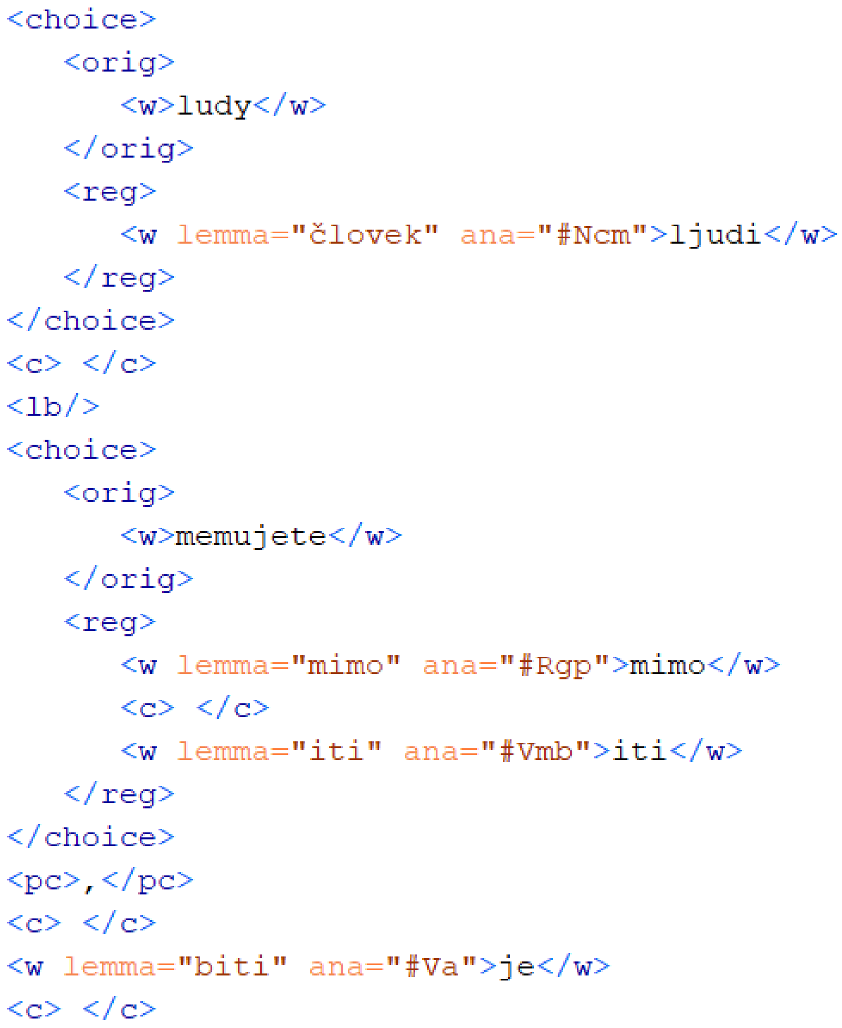
```
<s xml:id="ssj1.1.1">
    <pc ana="mte:Z" msd="UposTag=PUNCT" xml:id="ssj1.1.1.t1">"</pc>
    <w ana="mte:Pd-msg" lemma="tisti" xml:id="ssj1.1.1.t2">Tistega</w><c> </c>
    <w ana="mte:Ncmsg" lemma="večer" xml:id="ssj1.1.1.t3">večera</w><c> </c>
    <w ana="mte:Va-r1s-n" lemma="biti" xml:id="ssj1.1.1.t4">sem</w><c> </c>
    <w ana="mte:Rgp" lemma="preveč" xml:id="ssj1.1.1.t5">preveč</w><c> </c>
    <w ana="mte:Vmep-sm" lemma="popiti" xml:id="ssj1.1.1.t6">popil</w>
    <pc ana="mte:Z" xml:id="ssj1.1.1.t7">,</pc><c> </c>
    <w ana="mte:Vmep-sn" lemma="zgoditi" xml:id="ssj1.1.1.t8">zgodilo</w><c> </c>
    <w ana="mte:Px------y" lemma="se" xml:id="ssj1.1.1.t9">se</w><c> </c>
    <w ana="mte:Va-r3s-n" lemma="biti" xml:id="ssj1.1.1.t10">je</w><c> </c>
    <w ana="mte:Ncmsan" lemma="mesec" xml:id="ssj1.1.1.t11">mesec</w><c> </c>
    <w ana="mte:Ncmpg" lemma="dan" xml:id="ssj1.1.1.t12">dni</w><c> </c>
    <w ana="mte:Sl" lemma="po" xml:id="ssj1.1.1.t13">po</w><c> </c>
    <w ana="mte:Pd-nsl" lemma="ta" xml:id="ssj1.1.1.t14">tem</w>
    <pc ana="mte:Z" xml:id="ssj1.1.1.t15">,</pc><c> </c>
    <w ana="mte:Cs" lemma="ko" xml:id="ssj1.1.1.t16">ko</w><c> </c>
    <w ana="mte:Va-r1s-n" lemma="biti" xml:id="ssj1.1.1.t17">sem</w><c> </c>
    <w ana="mte:Vmep-sm" lemma="izvedeti" xml:id="ssj1.1.1.t18">izvedel</w>
    <pc ana="mte:Z" xml:id="ssj1.1.1.t19">,</pc><c> </c>
    <w ana="mte:Cs" lemma="da" xml:id="ssj1.1.1.t20">da</w><c> </c>
    <w ana="mte:Pp1-sa--y" lemma="jaz" xml:id="ssj1.1.1.t21">me</w><c> </c>
    <w ana="mte:Ncfsn" lemma="žena" xml:id="ssj1.1.1.t22">žena</w><c> </c>
    <w ana="mte:Vmpr3s" lemma="varati" xml:id="ssj1.1.1.t23">vara</w>
    <pc ana="mte:Z" xml:id="ssj1.1.1.t24">.</pc>
</s>
```

ssj500k v2.2: syntax



```xml
<linkGrp corresp="#ssj1.1.1" targFunc="head argument" type="UD-SYN">
  <link ana="ud-syn:punct" target="#ssj1.1.1.t6 #ssj1.1.1.t1"/>
  <link ana="ud-syn:det" target="#ssj1.1.1.t3 #ssj1.1.1.t2"/>
  <link ana="ud-syn:obl" target="#ssj1.1.1.t6 #ssj1.1.1.t3"/>
  <link ana="ud-syn:aux" target="#ssj1.1.1.t6 #ssj1.1.1.t4"/>
  <link ana="ud-syn:advmod" target="#ssj1.1.1.t6 #ssj1.1.1.t5"/>
  <link ana="ud-syn:root" target="#ssj1.1.1 #ssj1.1.1.t6"/>
  <link ana="ud-syn:punct" target="#ssj1.1.1.t8 #ssj1.1.1.t7"/>
  <link ana="ud-syn:parataxis" target="#ssj1.1.1.t6 #ssj1.1.1.t8"/>
  <link ana="ud-syn:expl" target="#ssj1.1.1.t8 #ssj1.1.1.t9"/>
  <link ana="ud-syn:aux" target="#ssj1.1.1.t8 #ssj1.1.1.t10"/>
  <link ana="ud-syn:obl" target="#ssj1.1.1.t8 #ssj1.1.1.t11"/>
  <link ana="ud-syn:nmod" target="#ssj1.1.1.t11 #ssj1.1.1.t12"/>
  <link ana="ud-syn:case" target="#ssj1.1.1.t14 #ssj1.1.1.t13"/>
  <link ana="ud-syn:nmod" target="#ssj1.1.1.t11 #ssj1.1.1.t14"/>
```

goo300k: Historical Slovene



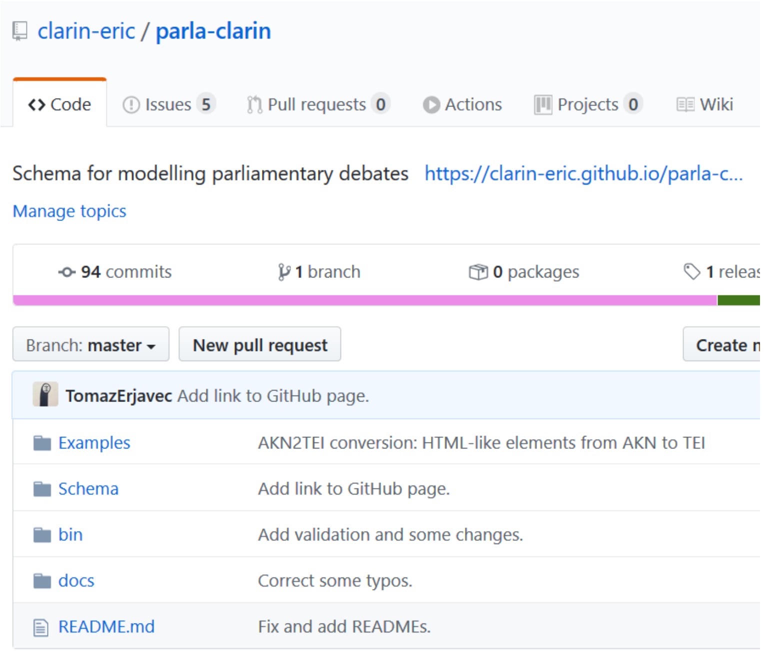
```
<choice>
   <orig>
      <w>ludy</w>
   </orig>
   <reg>
      <w lemma="človek" ana="#Ncm">ljudi</w>
   </reg>
</choice>
<c> </c>
<lb/>
<choice>
   <orig>
      <w>memujete</w>
   </orig>
   <reg>
      <w lemma="mimo" ana="#Rgp">mimo</w>
      <c> </c>
      <w lemma="iti" ana="#Vmb">iti</w>
   </reg>
</choice>
<pc>,</pc>
<c> </c>
<w lemma="biti" ana="#Va">je</w>
<c> </c>
```

Parla-CLARIN and siParl 2.0

- Parliamentary data: openly available and very useful for a number of disciplines
- Many countries have already made available corpora of parliamentary debates, but each one encoded differently
- CLARIN organised several events and initiatives to promote and study parliamentary corpora (PC)
- In 2019 CLARIN also organised a workshop and financed work to develop a common encoding schema for PCs
- Cooperation between DARIAH-SI (Andrej Pančur) and CLARIN.SI (Tomaž Erjavec):
  - TEI based Parla-CLARIN schema
  - siParl 2.0: first corpus encoded in Parla-CLARIN

Parla-CLARIN

- TEI based Parla-CLARIN schema: https://github.com/clarin-eric/parla-clarin
- ODD document + XML schema + example folders
- Readable documentation in HTML: https://clarin-eric.github.io/parla-clarin/



clarin-eric / parla-clarin

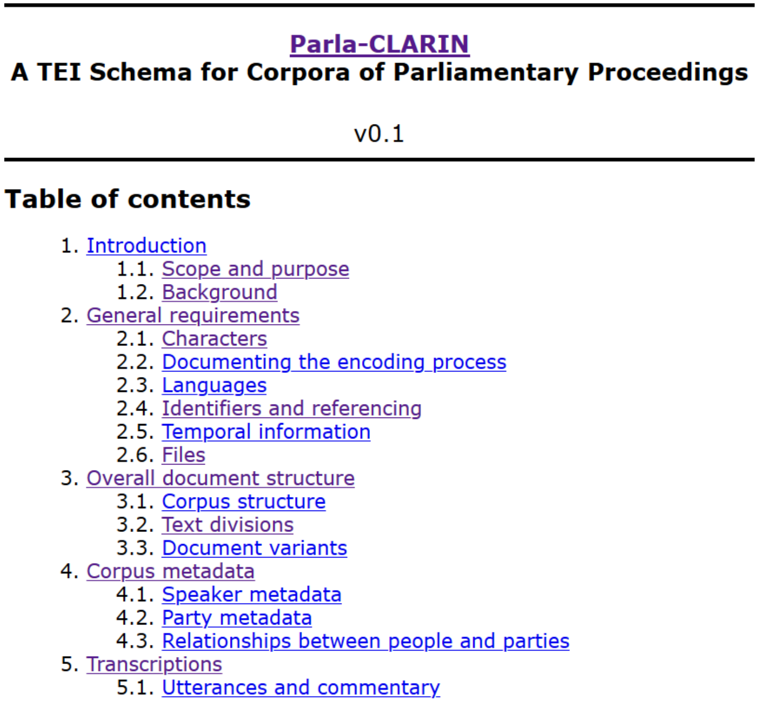
Schema for modelling parliamentary debates https://clarin-eric.github.io/parla-c...

Manage topics

94 commits | 1 branch | 0 packages | 1 releas

TomazErjavec Add link to GitHub page.

Examples	AKN2TEI conversion: HTML-like elements from AKN to TEI
Schema	Add link to GitHub page.
bin	Add validation and some changes.
docs	Correct some typos.
README.md	Fix and add READMEs.



Parla-CLARIN

A TEI Schema for Corpora of Parliamentary Proceedings

v0.1

Table of contents

1. Introduction
    1.1. Scope and purpose
    1.2. Background
2. General requirements
    2.1. Characters
    2.2. Documenting the encoding process
    2.3. Languages
    2.4. Identifiers and referencing
    2.5. Temporal information
    2.6. Files
3. Overall document structure
    3.1. Corpus structure
    3.2. Text divisions
    3.3. Document variants
4. Corpus metadata
    4.1. Speaker metadata
    4.2. Party metadata
    4.3. Relationships between people and parties
5. Transcriptions
    5.1. Utterances and commentary

siParl 2.0

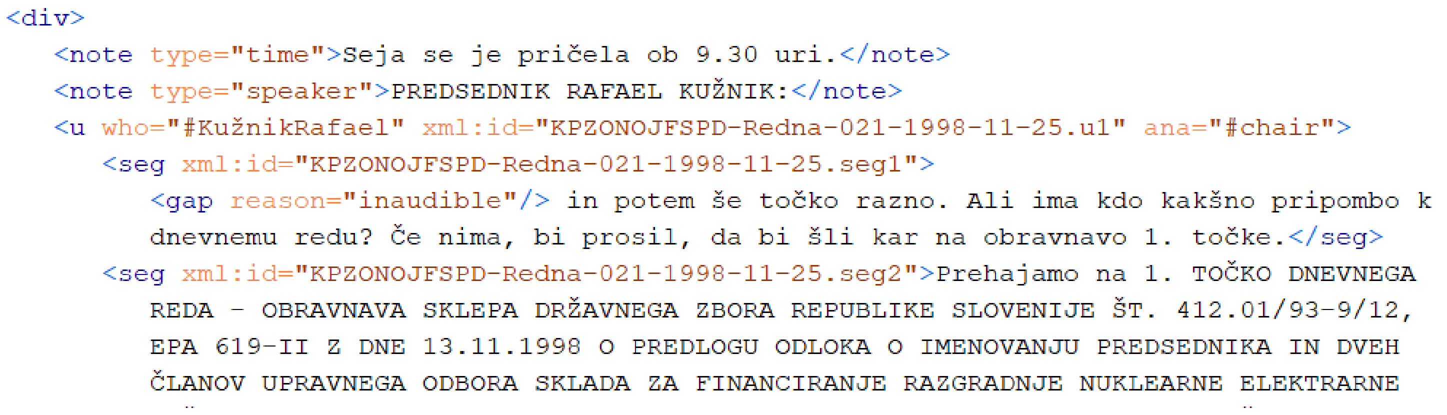
- First corpus encoded in Parla-CLARIN http://hdl.handle.net/11356/1300
- Slovene parliament 1990-2018
- Speaker meta-data
- Mandates, Parties, Committees, …
- Speeches
- CLARIN.SI version: linguistic annotation
- Available in the repository and through CLARIN.SI concordancers

siParl 2.0: metadata

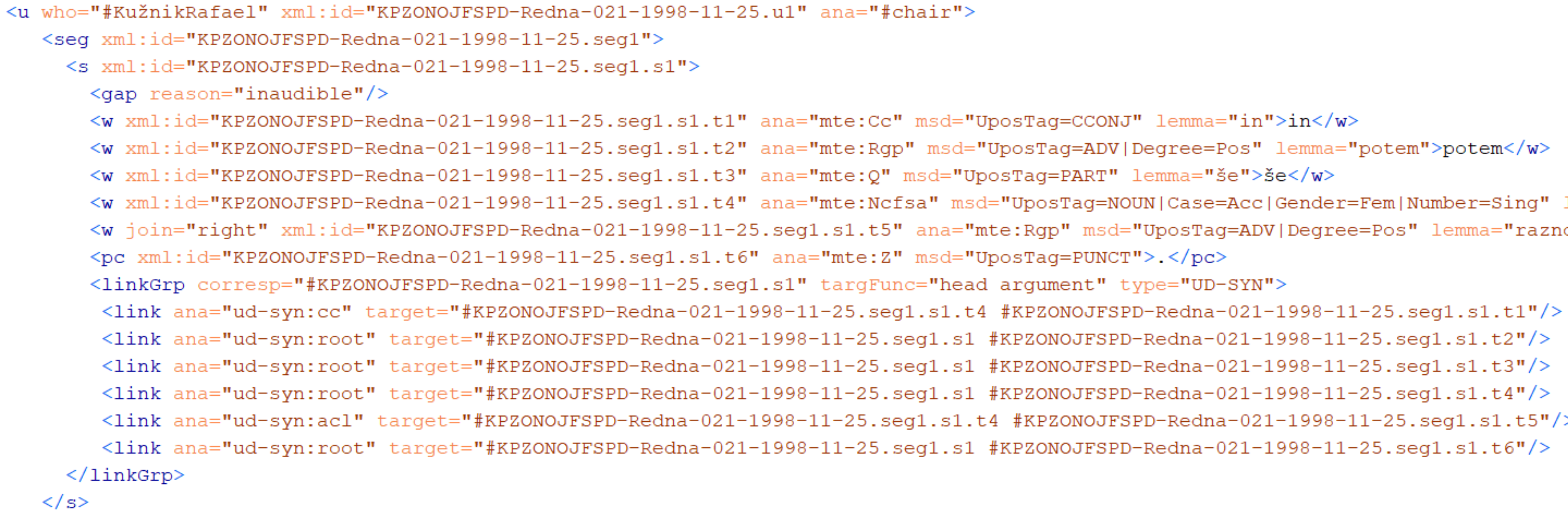
```
<org xml:id="party.Levica.1" role="political_party">
   <orgName full="yes" xml:lang="sl">Združena levica</orgName>
   <orgName full="yes" xml:lang="en">United Left</orgName>
   <orgName full="init">Levica</orgName>
   <event from="2014-03-01" to="2017-06-24">
      <label xml:lang="en">existence</label>
   </event>
   <idno type="wikimedia" xml:lang="sl">https://sl.wikipedia.org/wiki/Levica_(politi%C4%8Dna_stranka)</idno>
   <idno type="wikimedia" xml:lang="en">https://en.wikipedia.org/wiki/United_Left_(Slovenia)</idno>
</org>
```

```
<person xml:id="BavčarIgor">
   <persName>
      <surname>Bavčar</surname>
      <forename>Igor</forename>
   </persName>
   <sex value="M"/>
   <birth when="1955-11-28">
      <placeName ref="https://www.geonames.org/3196359">Ljubljana</placeName>
   </birth>
   <affiliation role="MP" ref="#DZ" from="1992-12-23" to="1996-11-27" ana="#DZ.1"/>
   <affiliation role="member" ref="#party.D" from="1992-12-23" to="1994-03-11" ana="#DZ.1"/>
   <affiliation role="member" ref="#party.LDS.2" from="1994-03-12" to="1996-11-27" ana="#DZ.1"/>
   <affiliation role="MP" ref="#DZ" from="1996-11-28" to="2000-10-26" ana="#DZ.2"/>
   <affiliation role="member" ref="#party.LDS.2" from="1996-11-28" to="2000-10-26" ana="#DZ.2"/>
   <affiliation role="MP" ref="#DZ" from="2000-10-27" to="2004-10-21" ana="#DZ.3"/>
   <affiliation role="member" ref="#party.LDS.2" from="2000-10-27" to="2004-10-21" ana="#DZ.3"/>
   <idno type="wikimedia" xml:lang="sl">https://sl.wikipedia.org/wiki/Igor_Bav%C4%8Dar</idno>
</person>
```

siParl: speeches



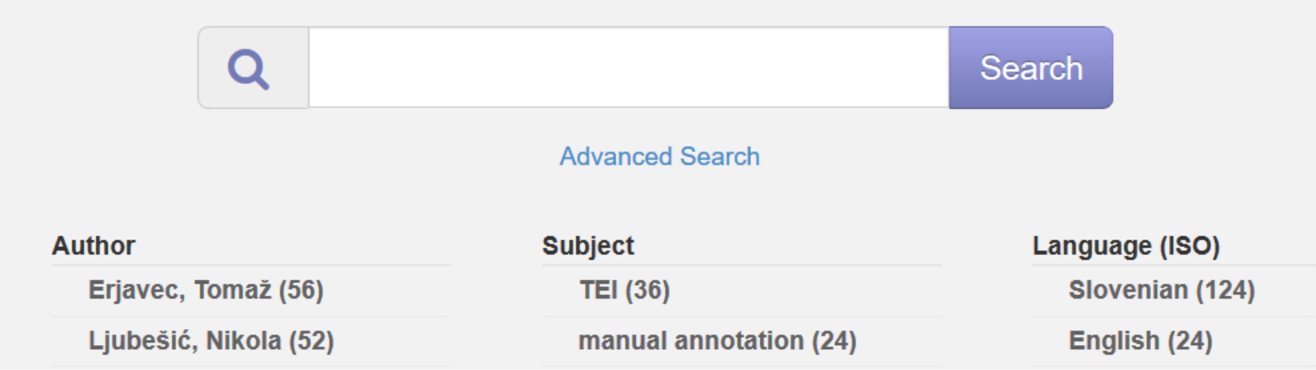
```
<div>
   <note type="time">Seja se je pričela ob 9.30 uri.</note>
   <note type="speaker">PREDSEDNIK RAFAEL KUŽNIK:</note>
   <u who="#KužnikRafael" xml:id="KPZONOJFSPD-Redna-021-1998-11-25.u1" ana="#chair">
      <seg xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1">
         <gap reason="inaudible"/> in potem še točko razno. Ali ima kdo kakšno pripombo k
         dnevnemu redu? Če nima, bi prosil, da bi šli kar na obravnavo 1. točke.</seg>
      <seg xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg2">Prehajamo na 1. TOČKO DNEVNEGA
         REDA - OBRAVNAVA SKLEPA DRŽAVNEGA ZBORA REPUBLIKE SLOVENIJE ŠT. 412.01/93-9/12,
         EPA 619-II Z DNE 13.11.1998 O PREDLOGU ODLOKA O IMENOVANJU PREDSEDNIKA IN DVEH
         ČLANOV UPRAVNEGA ODBORA SKLADA ZA FINANCIRANJE RAZGRADNJE NUKLEARNE ELEKTRARNE
```



```
<u who="#KužnikRafael" xml:id="KPZONOJFSPD-Redna-021-1998-11-25.u1" ana="#chair">
   <seg xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1">
     <s xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1">
       <gap reason="inaudible"/>
       <w xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t1" ana="mte:Cc" msd="UposTag=CCONJ" lemma="in">in</w>
       <w xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t2" ana="mte:Rgp" msd="UposTag=ADV|Degree=Pos" lemma="potem">potem</w>
       <w xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t3" ana="mte:Q" msd="UposTag=PART" lemma="še">še</w>
       <w xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t4" ana="mte:Ncfsa" msd="UposTag=NOUN|Case=Acc|Gender=Fem|Number=Sing" l
       <w join="right" xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t5" ana="mte:Rgp" msd="UposTag=ADV|Degree=Pos" lemma="razno
       <pc xml:id="KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t6" ana="mte:Z" msd="UposTag=PUNCT">.</pc>
       <linkGrp corresp="#KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1" targFunc="head argument" type="UD-SYN">
        <link ana="ud-syn:cc" target="#KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t4 #KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t1"/>
        <link ana="ud-syn:root" target="#KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1 #KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t2"/>
        <link ana="ud-syn:root" target="#KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1 #KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t3"/>
        <link ana="ud-syn:root" target="#KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1 #KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t4"/>
        <link ana="ud-syn:acl" target="#KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t4 #KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t5"/>
        <link ana="ud-syn:root" target="#KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1 #KPZONOJFSPD-Redna-021-1998-11-25.seg1.s1.t6"/>
     </linkGrp>
   </s>
```

Conclusions

- TEI is a large collections of elements and attributes for annotating various types of texts and their analyses
- Documented, maintained, with lots of support tools and a large user community
- Relatively popular in Slovenia in Digital Humanities, Corpus and Computational Linguistics
- Many CC TEI language resources (with down-conversions) are available in the CLARIN.SI repository



Author	Subject	Language (ISO)
Erjavec, Tomaž (56)	TEI (36)	Slovenian (124)
Ljubešić, Nikola (52)	manual annotation (24)	English (24)