

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 \(en\)](#)



RESEARCH & INNOVATION

Infrastructures

[European Commission](#) > [Research & Innovation](#) > [Research infrastructures](#) > [ESFRI](#)



- HOME
- WHAT ARE RIs ?
- MAPS of RIs
- THE EUROPEAN LANDSCAPE
- EU FINANCIAL SUPPORT
- ERIC-LEGAL FRAMEWORK
- SYNERGIES - EU INITIATIVES
- INTERNATIONAL COOPERATION



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](#)



Highlights



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





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



- 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


[About ▼](#)
[Participants](#)
[Services](#)
[Knowledge Base ▼](#)
[Funding](#)
[Events](#)
[News](#)
[Contact](#)

[Applications](#)
[Intranet login](#)

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
Common Language Resources and
Technology Infrastructure

CLARIN Funding for Virtual Events



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 “Jezikovnotehnoški 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)



- 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)

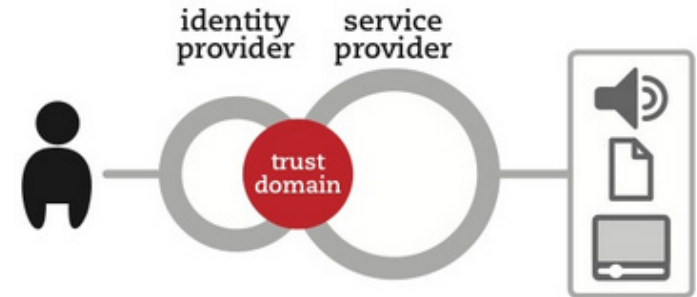
CLARIN.SI Cooperation

- 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

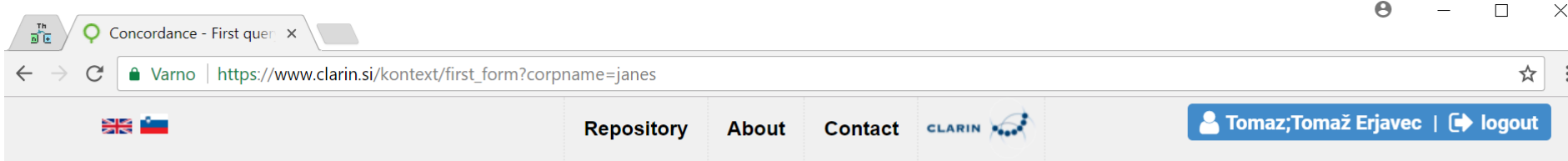


- Infrastructure for authentication and authorisation (AAI)
- Single Sign-On: separation between the Identity Provider and Service Provider
- As opposed to standard web login we here know the identity of the user
- Identity provider federations: EduGain
- Slovene users can, via EduGain, access most CLARIN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



[Query](#) [Corpora](#) [Save](#) [Concordance](#) [Filter](#) [Frequency](#) [Collocations](#) [View](#) [Help](#)

Corpus: [Janes \(družbena omrežja\)](#)

Search in the corpus

Corpus:

Query Type: ?

Query: [keyboard](#) | [recent queries](#)

► Specify context

► Specify query according to the meta-information

Concordances

The screenshot shows the KonText web interface in a browser window. The address bar shows the URL: 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.... The page has a header with navigation links: Repository, About, Contact, and a user profile for Tomaz;Tomaž Erjavec with a logout button. The main content area shows the 'kon text' logo and navigation links: Query, Corpora, Save, Concordance, Filter, Frequency, Collocations, View, Help. Below this, it indicates the corpus is 'Janes (družbena omrežja)' and the query is 'krava' with 6,958 hits. A green bar displays statistics: Hits: 6,958 | i.p.m.: 27.51 (related to the whole "janes") | ARF: 3,261.05 | Result is sorted. Below the statistics, there is a 'Line selection' dropdown set to 'simple' and a 'Display options' link. The main table displays concordance results with columns for source text, the query word 'krava', and the target text. The results are sorted by frequency, with 'krava' appearing in various contexts, often related to 'krave' or 'krav'.

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

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 navrh 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](#)

Page [Next >](#)

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](#)

Page [Next >](#)

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
člen	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)

The screenshot displays the WebAnno (CLARIN-DE) web interface. The browser address bar shows www.clarin.si/webanno/curation.html?6. The interface has a red header bar with the title "Curation" and a "WebAnno | Home" link. Below the header is a navigation bar with tabs for "Document", "Page", "Script", "Help", and "Workflow". The "Document" tab is active, showing a toolbar with icons for "Open", "Re-create", "Merge", "Prev.", "Next", "Export", and "Settings". The "Page" tab shows "Page 7" with navigation buttons "First", "Prev.", "Go to", "Next", and "Last". The "Script" tab shows "LTR/RTL". The "Help" tab shows "Guidelines". The "Workflow" tab shows "Finish".

The main content area is divided into three sections: "Sentences", "Annotation", and "Actions". The "Sentences" section on the left lists sentences 1 through 18, with sentence 7 highlighted. The "Annotation" section displays the text of sentence 7: "Avtorja pravita , da parameter Lndim zajema vpliv učinkov ostenja (angl. boundary effects) , zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .". Above the text, there are two annotations: "2TermSlv" and "1TermEng", connected by a red arrow labeled "(kas. Translation)". The "Actions" section on the right shows a dropdown menu for "Layer" set to "kas.BiTerm", a checkbox for "Forward annotation ?" which is unchecked, and a message "No annotation selected!".

At the bottom of the interface, there is a status bar showing "Technische Universität Darmstadt -- Computer Science Department -- WebAnno -- 3.2.2 (2017-07-18 23:57:48, build 61cea1f0f7eda3b57d7c548d0577f166ac2830ce)". The Windows taskbar at the very bottom shows the system clock as 19:13 on 07/11/2017.

ReLDI automatic text annotation

- Web form
- API

Query

Tagger **Lexicon**

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

PROCESS **CLEAR**

or

File Browse... No file selected. **REMOVE**

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