The Slovene CLARIN Infrastructure

Tomaž Erjavec

Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia

The objective of the Slovene research infrastructure CLARIN.SI is to facilitate research in the humanities and social sciences by offering researchers authorised access to its platform, which will integrate Slovene language resources and advanced tools for processing of Slovene. The Slovene CLARIN is based at the Jožef Stefan Institute and has received initial funding in 2013-2014. In 2014 CLARIN.SI has been established as a consortium of all the main institutions involved in language technologies and linguistic research. Currently, CLARIN.SI offers access to the platform of the project "Communication in Slovene" and to a number of Slovene corpora via a web-based concordancer and has set up its web pages and a test version of the LINDAT repository.

History of Slovene CLARIN

- **2011**: Slovene Government adopts the Research infrastructure Roadmap 2011-2020, which includes CLARIN, as well as DARIAH and CESSDA;
- **2012**: DARIAH and CESSDA establish Slovene infrastructures SI-DIH and ADP;
- **October 2013**: surprise (minimal) funding for CLARIN @ Jožef Stefan Institute;
- **February 2014:** Slovenian Language Technologies Society publishes a call for the creation of the consortium of Slovenian CLARIN.SI;
- **March 2014**; transfer of the www-slovenscina.eu portal to JSI infrastructure
- **June 2014**: establishment of the CLARIN.SI Consortium;
- **July 2014**: extraordinary elections in Slovenia, meaning that the signing of the agreement for Slovenia to join CLARIN ERIC is, once again, postponed.

CLARIN.SI Consortium

Universities

University of Ljubljana

At the University of Ljubljana, corpus linguistics and language technology research is carried out at the Centre for Language Technologies, at the Faculty of Arts (primarily the Department of Translation), Faculty of Social Sciences (Research Centre for the Terminology of Social Sciences and Journalism), Faculty of Electrical Engineering (Laboratory of Artificial Perception, Systems and Cybernetics) and the Faculty of Computer and Information Science (Laboratory for Cognitive Modeling).

University of Maribor

Research on language and speech technologies at the University of Maribor is undertaken mainly at the Faculty of Electrical Engineering and Computer Sciences, in the scope of the Institute for Electronics and Telecommunications and the Institute for Computer Science (Laboratory for Heterogenous Computer systems).

University of Primorska

Language technology research and corpus linguistics are undertaken at the Faculty for Mathematics, Natural Sciences and Information Technologies, mainly at the Department of Information Scineces and Technologies.

Companies

Alpineon, d.o.o.

The development of hardware and software in the field of language and speech technologies: speech recognition and synthesis, machine translation, voice portals, SMS and e-mail readers.

Amebis, d.o.o., Kamnik

Language technology software: Slovenian spell and syntax checkers, machine translations, speech synthesis, corpora, online dictionaries, virtual agents, etc.

Institutes

Jožef Stefan Institute

At JSI several departments (E3, E8, E9) are involved in language technology research, working in the fields of linguistic annotation of texts, compiling language corpora and other Slovene language resources, text mining and text analytics, machine translation, speech synthesis, etc.

Institute of Contemporary History

The institute is, together with the SRC SASA, in charge of the SI-DIH portal, the Slovenian branch of the European research infrastructure for the Humanities DARIAH.

Scientific and Research Centre of the Slovenian Academy of Sciences and Arts

At SRC SASA, the Fran Ramovš Institute for Slovenian Language is the national centre for systematic monitoring and description of Slovene language materials. The results of this research are dictionaries, collections and reference works for Slovene language and scientific papers.

Societies

Slovenian Language Technologies Society

The Society was founded in 1998 and joins people working on language technologies from the scientific, educational or user perspectives. Its activities are aimed at promoting the development of language technologies for the Slovenian language.

Domestic Research Society

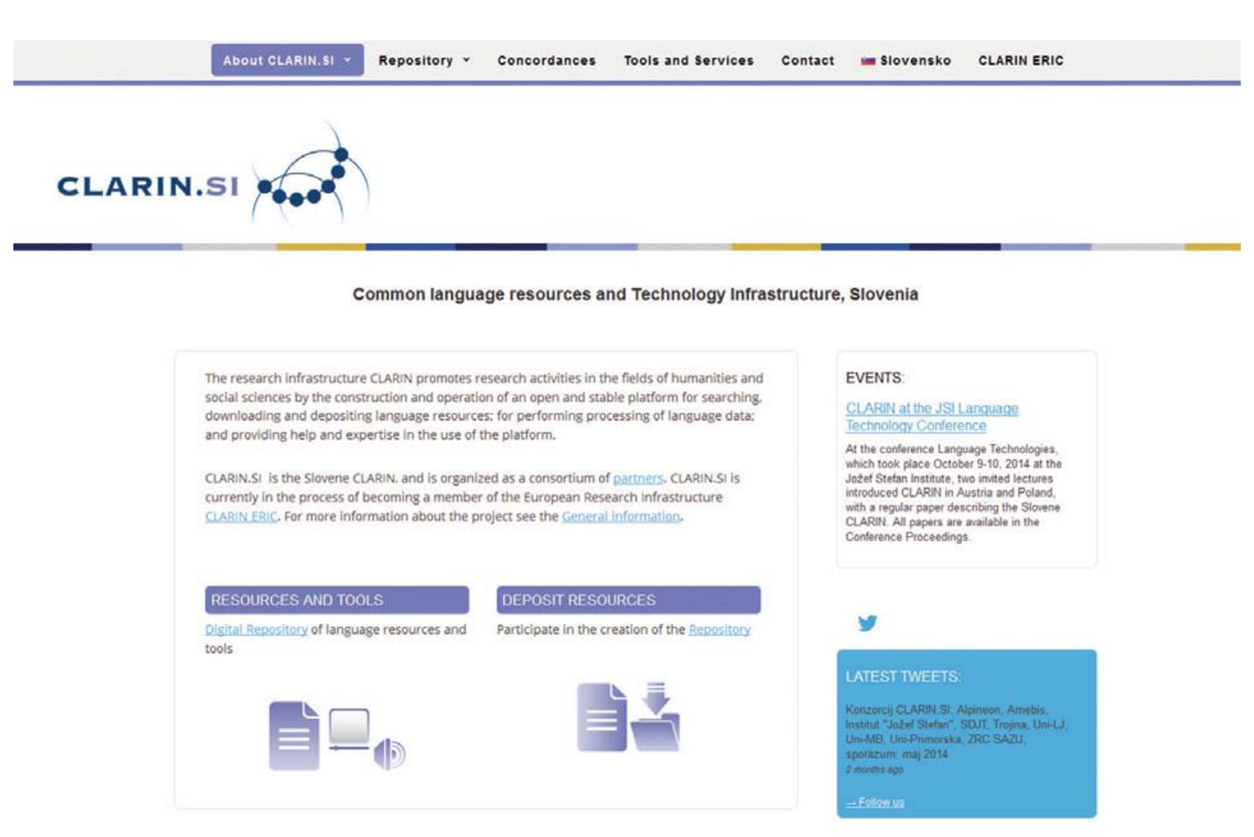
Domestic Research Society manages the first free online dictionary of the Slovene spoken language, "The Unleashed Tongue", developed in 2004.

Trojina, Institute for Applied Slovene Studies

Trojina is undertakes projects aimed at modern, targeted linguistic research and at increasing the confidence of speakers in public and private use of the Slovene language.

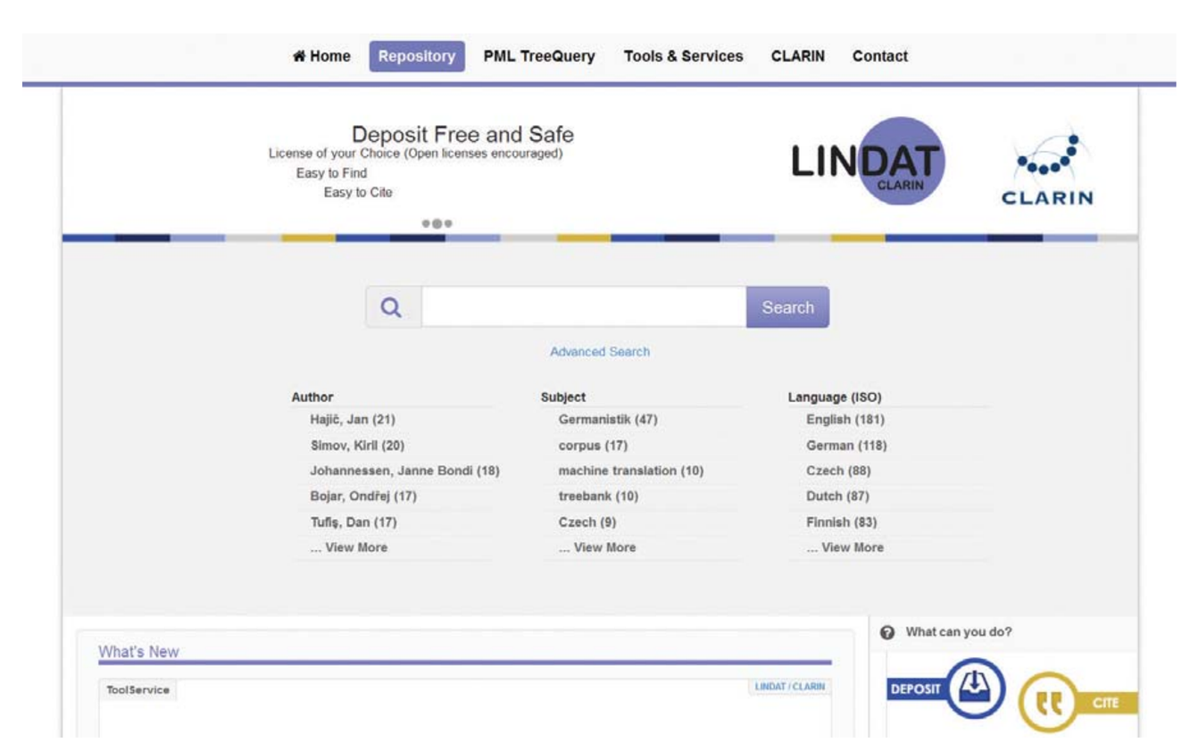
Web page

- Web pages introduce the infrastructure and consortium partners
- Information in Slovene and English
- Links to the repository and services



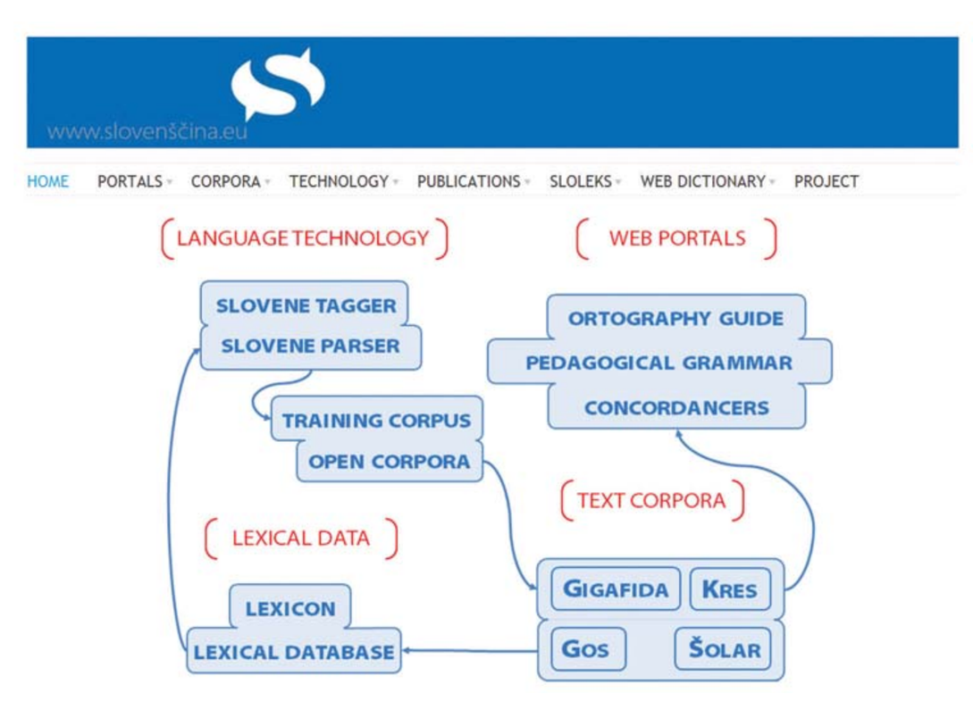
LINDAT Repository

- CLARIN.SI uses the Czech UFAL CLARIN LINDAT repository platform
- Based on DSpace, available on GitHub
- Supports handles and AAI
- Currently in testing



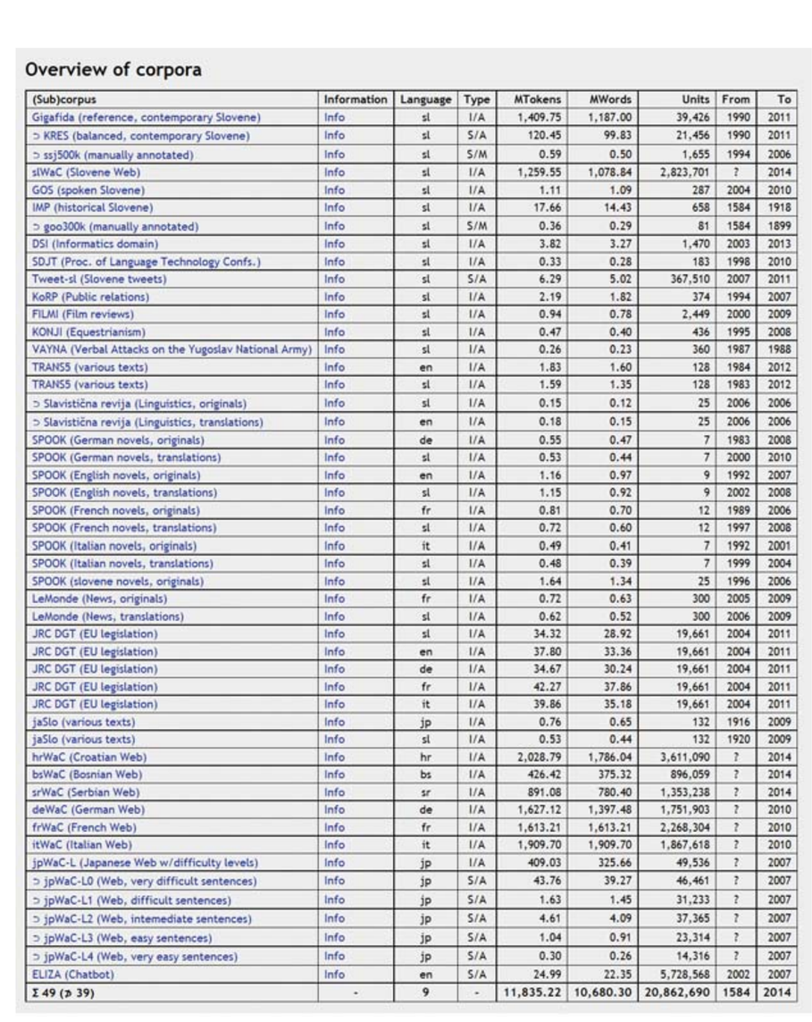
www.slovenscina.eu

- CLARIN.SI hosts the platform of the project »Communication in Slovene«
- Web concordancers to reference corpora of Slovene
- Web services for tagging, lemmatising and parsing of Slovene
- Downloadable training corpora, morphological lexicon and lexical database



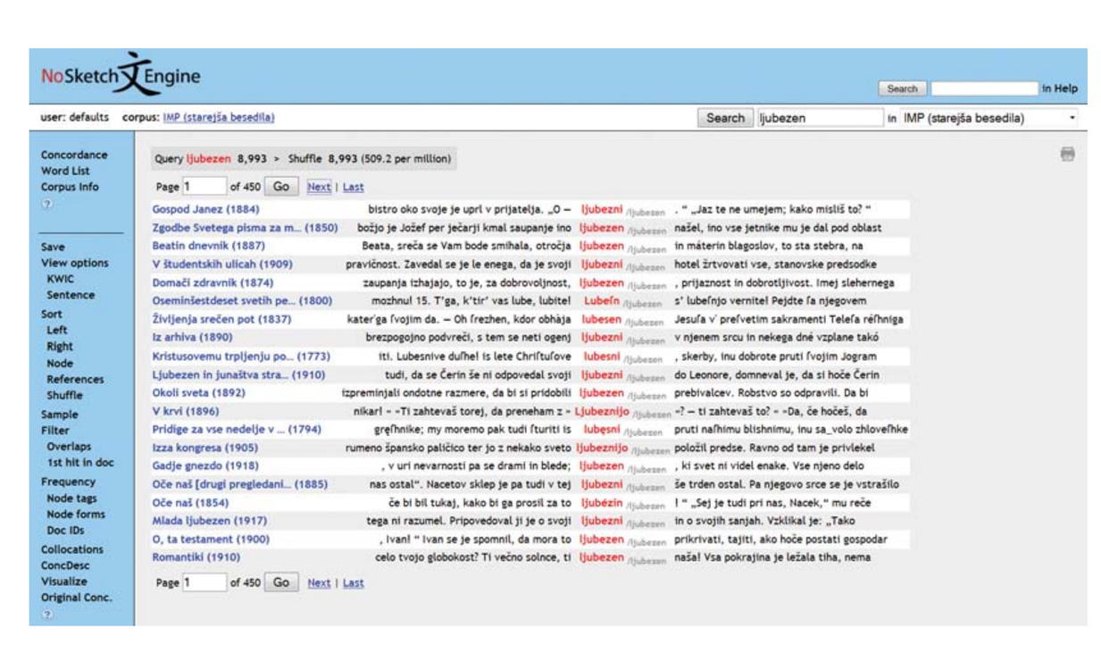
Searching Slovene Corpora

- Freely available concordancing over many corpora
- Reference and specialised monolingual and parallel corpora
- Linguistic annotations: normalised word-forms, PoS tags, lemmas
- Corpora mounted on noSketchEngine and CUWI concordancers
- Powerful CWB queries and fast processing
- Localised interfaces



Overview of corpora

(Sub)corpus	Information	Language	Type	MTokens	MWords	Units	From	To
Gigafida (reference, contemporary Slovene)	Info	sl	I/A	1,409.75	1,187.00	39,426	1990	2011
⊃ KRES (balanced, contemporary Slovene)	Info	sl	S/A	120.45	99.83	21,456	1990	2011
⊃ ssj500k (manually annotated)	Info	sl	S/M	0.59	0.50	1,655	1994	2006
slWaC (Slovene Web)	Info	sl	I/A	1,259.55	1,078.84	2,823,701	?	2014
GOS (spoken Slovene)	Info	sl	I/A	1.11	1.09	287	2004	2010
IMP (historical Slovene)	Info	sl	I/A	17.66	14.43	658	1584	1918
⊃ goo300k (manually annotated)	Info	sl	S/M	0.36	0.29	81	1584	1899
DSI (Informatics domain)	Info	sl	I/A	3.82	3.27	1,470	2003	2013
SDJT (Proc. of Language Technology Confs.)	Info	sl	I/A	0.33	0.28	183	1998	2010
Tweet-sl (Slovene tweets)	Info	sl	S/A	6.29	5.02	367,510	2007	2011
KoRP (Public relations)	Info	sl	I/A	2.19	1.82	374	1994	2007
FILMI (Film reviews)	Info	sl	I/A	0.94	0.78	2,449	2000	2009
KONJI (Equestrianism)	Info	sl	I/A	0.47	0.40	436	1995	2008
VAYNA (Verbal Attacks on the Yugoslav National Army)	Info	sl	I/A	0.26	0.23	360	1987	1988
TRANS5 (various texts)	Info	en	I/A	1.83	1.60	128	1984	2012
TRANS5 (various texts)	Info	sl	I/A	1.59	1.35	128	1983	2012
⊃ Slavistična revija (Linguistics, originals)	Info	sl	I/A	0.15	0.12	25	2006	2006
⊃ Slavistična revija (Linguistics, translations)	Info	en	I/A	0.18	0.15	25	2006	2006
SPOOK (German novels, originals)	Info	de	I/A	0.55	0.47	7	1983	2008
SPOOK (German novels, translations)	Info	sl	I/A	0.53	0.44	7	2000	2010
SPOOK (English novels, originals)	Info	en	I/A	1.16	0.97	9	1992	2007
SPOOK (English novels, translations)	Info	sl	I/A	1.15	0.92	9	2002	2008
SPOOK (French novels, originals)	Info	fr	I/A	0.81	0.70	12	1989	2006
SPOOK (French novels, translations)	Info	sl	I/A	0.72	0.60	12	1997	2008
SPOOK (Italian novels, originals)	Info	it	I/A	0.49	0.41	7	1992	2001
SPOOK (Italian novels, translations)	Info	sl	I/A	0.48	0.39	7	1999	2004
SPOOK (slovene novels, originals)	Info	sl	I/A	1.64	1.34	25	1996	2006
LeMonde (News, originals)	Info	fr	I/A	0.72	0.63	300	2005	2009
LeMonde (News, translations)	Info	sl	I/A	0.62	0.52	300	2006	2009
JRC DGT (EU legislation)	Info	sl	I/A	34.32	28.92	19,661	2004	2011
JRC DGT (EU legislation)	Info	en	I/A	37.80	33.36	19,661	2004	2011
JRC DGT (EU legislation)	Info	de	I/A	34.67	30.24	19,661	2004	2011
JRC DGT (EU legislation)	Info	fr	I/A	42.27	37.86	19,661	2004	2011
JRC DGT (EU legislation)	Info	it	I/A	39.86	35.18	19,661	2004	2011
jaSlo (various texts)	Info	jp	I/A	0.76	0.65	132	1916	2009
jaSlo (various texts)	Info	sl	I/A	0.53	0.44	132	1920	2009
hrWaC (Croatian Web)	Info	hr	I/A	2,028.79	1,786.04	3,611,090	?	2014
bsWaC (Bosnian Web)	Info	bs	I/A	426.42	375.32	896,059	?	2014
srWaC (Serbian Web)	Info	sr	I/A	891.08	780.40	1,353,238	?	2014
deWaC (German Web)	Info	de	I/A	1,627.12	1,397.48	1,751,903	?	2010
frWaC (French Web)	Info	fr	I/A	1,613.21	1,613.21	2,268,304	?	2010
itWaC (Italian Web)	Info	it	I/A	1,909.70	1,909.70	1,867,618	?	2010
jpWaC-L (Japanese Web w/difficulty levels)	Info	jp	I/A	409.03	325.66	49,536	?	2007
⊃ jpWaC-L0 (Web, very difficult sentences)	Info	jp	S/A	43.76	39.27	46,461	?	2007
⊃ jpWaC-L1 (Web, difficult sentences)	Info	jp	S/A	1.63	1.45	31,233	?	2007
⊃ jpWaC-L2 (Web, intemediate sentences)	Info	jp	S/A	4.61	4.09	37,365	?	2007
⊃ jpWaC-L3 (Web, easy sentences)	Info	jp	S/A	1.04	0.91	23,314	?	2007
⊃ jpWaC-L4 (Web, very easy sentences)	Info	jp	S/A	0.30	0.26	14,316	?	2007
ELIZA (Chatbot)	Info	en	S/A	24.99	22.35	5,728,568	2002	2007
Σ 49 (⊅ 39)	-	9	-	11,835.22	10,680.30	20,862,690	1584	2014



www.clarin.si

ACKNOWLEDGEMENT: CLARIN.SI is supported by the Slovenian Ministry of Education, Science and Sport