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;

CLARIN.SI Consortium								
Universities		Institutes						
University of Ljubljana Univerza <i>v Ljubljani</i>	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 Elec- trical Engineering (Laboratory of Artificial Perception, Systems	Jožef Stefan Institute Institut "Jožef Stefan" Ljubljana, Slovenija	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.					
University of Maribor	and Cybernetics) and the Faculty of Computer and Information Science (Laboratory for Cognitive Modeling). Research on language and speech technologies at the University	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 infra- structure for the Humanities DARIAH.					
Univerza v Mariboru	of Maribor is undertaken mainly at the Faculty of Electrical Engi- neering and Computer Sciences, in the scope of the Institute for Electronics and Telecommunications and the Institute for Com- puter Science (Laboratory for Heterogenous Computer systems).	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 lan-					
University of Primorska	Language technology research and corpus linguistics are under- taken at the Faculty for Mathematics, Natural Sciences and In- formation Technologies, mainly at the Department of Information Scineces and Technologies.	SLOVENSE AKADENUE SLOVENSE AKADENUE	guage and scientific papers.					

- 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.

Companies

Kamnik

Societies

Trojina, Institute for Applied

Slovene Studies

trojína

The development of hardware and software in the field of language and Alpineon, d.o.o. **Slovenian Language** The Society was founded in 1998 and joins people working on speech technologies: speech recognition and synthesis, machine trans-**Technologies Society** language technologies from the scientific, educational or user alpineon) lation, voice portals, SMS and e-mail readers. perspectives. Its activities are aimed at promoting the develop-SDJT ment of language technologies for the Slovenian language. Amebis, d.o.o., Language technology software: Slovenian spell and syntax checkers, Domestic Research Society manages the first free online diction-**Domestic Research Society** machine translations, speech synthesis, corpora, online dictionaries, ary of the Slovene spoken language, "The Unleashed Tongue", virtual agents, etc. **AMEBIS** developed in 2004.

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

Trojina is undertakes projects aimed at modern, targeted linguistic research and at increasing the confidence of speakers in pub-

lic and private use of the Slovene language.

- 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



	Uni-MB, Uni-Primorska, ZRC SAZU, sporazum: maj 2014 2 movims apo	
	Follow us	

What's New	What can you do?
ToolService	

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

b)corpus	Information	Language	Туре	MTokens	MWords	Units	From	To
afida (reference, contemporary Slovene)	Info	sl	I/A	1,409.75	1,187.00	39,426	1990	2011
RES (balanced, contemporary Slovene)	Info	sl	S/A	120.45	99.83	21,456	1990	2011
sj500k (manually annotated)	Info	sl	S/M	0.59	0.50	1.655	1994	2006
aC (Slovene Web)	Info	sl	I/A	1.259.55	1.078.84	2.823.701	2	2014
(spoken Slovene)	Info	sl	I/A	1.11	1.09	287	2004	2010
(historical Slovene)	Info	sl	I/A	17.66	14.43	658	1584	1918
00300k (manually annotated)	Info	sl	S/M	0.36	0.29	81	1584	1899
(Informatics domain)	Info	sl	I/A	3.82	3.27	1,470	2003	2013
T (Proc. of Language Technology Confs.)	Info	sl	I/A	0.33	0.28	183	1998	2010
eet-sl (Slovene tweets)	Info	sl	S/A	6.29	5.02	367,510	2007	2011
P (Public relations)	Info	sl	I/A	2.19	1.82	374	1994	2007
Al (Film reviews)	Info	sl	I/A	0.94	0.78	2,449	2000	2009
UI (Equestrianism)	Info	sl	1/A	0.47	0.40	436	1995	2008
(NA (Verbal Attacks on the Yugoslav National Army)	Info	sl	I/A	0.4/	0.40	360	1987	1988
NS5 (various texts)	info	en	1/A	1.83	1.60	128	1984	2012
NS5 (various texts)	Info	en sl	I/A	1.63	1.35	120	1983	2012
	Info	ડા ડા	1/A	0.15	0.12	25	2006	2006
lavistična revija (Linguistics, originals)			1/A	0.15	0.12	25	2006	2006
lavistična revija (Linguistics, translations)	Info	en						
OK (German novels, originals)	Info	de	I/A	0.55	0.47	7	1983	2008
OK (German novels, translations)	Info	sl	I/A	0.53	0.44	7	2000	2010
IOK (English novels, originals)	Info	en	I/A	1.16	0.97	9	1992	2007
OK (English novels, translations)	Info	sl	I/A	1.15	0.92	9	2002	2008
OK (French novels, originals)	Info	fr	I/A	0.81	0.70	12	1989	2006
IOK (French novels, translations)	Info	sl	I/A	0.72	0.60	12	1997	2008
IOK (Italian novels, originals)	Info	ít	I/A	0.49	0.41	7	1992	2001
IOK (Italian novels, translations)	Info	sl	I/A	0.48	0.39	7	1999	2004
IOK (slovene novels, originals)	Info	sl	I/A	1.64	1.34	25	1996	2006
londe (News, originals)	Info	fr	I/A	0.72	0.63	300	2005	2009
londe (News, translations)	Info	sl	I/A	0.62	0.52	300	2006	2009
DGT (EU legislation)	Info	si	I/A	34.32	28.92	19,661	2004	2011
DGT (EU legislation)	Info	en	I/A	37.80	33.36	19,661	2004	2011
DGT (EU legislation)	Info	de	1/A	34.67	30.24	19,661	2004	2011
DGT (EU legislation)	Info	fr	I/A	42.27	37.86	19,661	2004	2011
DGT (EU legislation)	Info	ít	I/A	39.86	35.18	19,661	2004	2011
io (various texts)	Info	jp	I/A	0.76	0.65	132	1916	2009
o (various texts)	Info	sl	I/A	0.53	0.44	132	1920	2009
/aC (Croatian Web)	Info	hr	I/A	2,028.79	1,786.04	3,611,090	7	2014
/aC (Bosnian Web)	Info	bs	I/A	426.42	375.32	896,059	?	2014
aC (Serbian Web)	Info	sr	I/A	891.08	780.40	1,353,238	?	2014
VaC (German Web)	Info	de	I/A	1,627.12	1,397.48	1,751,903	7	2010
aC (French Web)	Info	fr	I/A	1,613.21	1,613.21	2,268,304	2	2010
aC (Italian Web)	Info	it	I/A	1,909.70	1,909.70	1,867,618	2	2010
/aC-L (Japanese Web w/difficulty levels)	Info	jp	I/A	409.03	325.66	49,536	?	2007
pWaC-L0 (Web, very difficult sentences)	Info	jp	S/A	43.76	39.27	46,461	?	2007
pWaC-L1 (Web, difficult sentences)	Info	jp	S/A	1.63	1.45	31,233	?	2007
pWaC-L2 (Web, internediate sentences)	Info	jp	S/A	4.61	4.09	37,365	7	2007
pWaC-L3 (Web, easy sentences)	Info	jp	S/A	1.04	0.91	23,314	?	2007
pWaC-L4 (Web, very easy sentences)	Info	jp	S/A	0.30	0.26	14,316	?	2007
IA (Chatbot)	Info	en	S/A	24.99	22.35	5,728,568	2002	2007
n (niemon)		en	D/A	24.99	10,680.30	20,862,690	2002	2007



s page gives entry points to the web-based concordancer and corpus analysis tool NoSketch Engine installed on the Natural Lan server of the Dept. of Knowledge Technolgies at the Jožef Stefan Institute. The server offers a number of corpora, the majority in Sloven ut also other languages (esp. English and Japanese). The corpora have been produced at JSI or in projects where JSI was a partner. Most of the corpora are available for web-based exploration without restrictions, excecpt to acknowledge their use by citing nl.ijs.si, and, in case of automatic access, of not abusing the serve

NoSketch Engine entry points

 All nl.ijs.si public corpora · Reference: contemporary, historical and spoken Sloven • Parallel: English, Japanese, French Specialised: terminologies, social web, recent history
Web: Slovene, Croatian, Serbian, Bosnian, German, French, Italian, Japanese • SPOOK: parallel en, de, fr, it-sl (access restricted) • ELIZA: dialogues with the famous chatbot (access restricted)

user: defaults co	rpus: IMP (starejša besedila)			Search	ljubezen	10	MP (starejša be:
user: dereutits ed	Pest inter (score) se oesecuter			Search	IJubezen	in i	Mr (starejsa be
Concordance Word List	Query ljubezen 8,993 > Shuffle 8,9						
Corpus Info	Page 1 of 450 Go Next 1	Last					
	Gospod Janez (1884)	bistro oko svoje je uprl v prijatelja. "O -	· gase series	. " "Jaz te ne u	mejem; kako misliš t	0? "	
	Zgodbe Svetega pisma za m (1850)	božjo je Jožef per ječarji kmal saupanje ino		našel, ino vse je	etnike mu je dal pod e	oblast	
Save	Beatin dnevnik (1887)	Beata, sreča se Vam bode smihala, otročja	ljubezen /ljubezen	in máterin blag	oslov, to sta stebra, i	na	
View options	V študentskih ulicah (1909)	pravičnost. Zavedal se je le enega, da je svoji	ljubezni /jubezen	hotel žrtvovati	vse, stanovske preds	odke	
KWIC	Domači zdravnik (1874)	zaupanja izhajajo, to je, za dobrovoljnost,	ljubezen Ajubezen	, prijaznost in d	obrotljivost. Imej sle	ehernega	
Sentence	Oseminšestdeset svetih pe (1800)	mozhnul 15. T'ga, k'tir' vas lube, lubite!	Lubein Ajubezen	s' lubeſnjo vern	itel Pejdte la njegovi	em	
Sort	Življenja srečen pot (1837)	kater'ga ívojim da Oh írezhen, kdor obhaja	lubesen /jubesen	Jesula v' prefve	tim sakramenti Telef	a réfhniga	
Left	Iz arhiva (1890)	brezpogojno podvreči, s tem se neti ogenj	ljubezni Ajubezen	v njenem srcu i	n nekega dné vzplane	takó	
Right Node	Kristusovemu trpljenju po (1773)	iti. Lubesnive dufhe! is lete Chriftufove	lubesni /liobezen	, skerby, inu do	brote pruti ívojim Jo	gram	
References	Ljubezen in junaštva stra (1910)	tudi, da se Čerin še ni odpovedal svoji	ljubezni Autom	do Leonore, dor	nneval je, da si hoče	Čerin	
Shuffle	Okoli sveta (1892)	izpreminjali ondotne razmere, da bi si pridobili		prebivalcev. Ro	bstvo so odpravili. Da	s bi	
Sample	V krvi (1896)	nikariTi zahtevaš torej, da preneham z - I		-? - ti zahtevaš	š to?Da, če hočeš,	da	
Filter	Pridige za vse nedelje v (1794)	grefhnike; my moremo pak tudi fturiti is			lishnimu, inu sa_volo	zhlovefhke	
Overlaps	Izza kongresa (1905)	rumeno špansko paličico ter jo z nekako sveto	ljubeznijo marte	položil predse.	Ravno od tam je privi	lekel	
1st hit in doc	Gadje gnezdo (1918)	, v uri nevarnosti pa se drami in blede;	ljubezen diskeren	, ki svet ni vide	l enake. Vse njeno de	elo	
Frequency	Oče naš [drugi pregledani (1885)	nas ostal". Nacetov sklep je pa tudi v tej		še trden ostal.	Pa njegovo srce se je	vstrašilo	
Node tags	Oče naš (1854)	če bi bil tukaj, kako bi ga prosil za to		I " "Sej je tudi j	pri nas, Nacek," mu r	eče	
Node forms	Mlada ljubezen (1917)	tega ni razumel. Pripovedoval ji je o svoji		in o svojih sanja	ah. Vzklikal je: "Tako		
Doc IDs	O, ta testament (1900)	, Ivan! " Ivan se je spomnil, da mora to		prikrivati, tajit	I, ako hoče postati go	spodar	
Collocations	Romantiki (1910)	celo tvojo globokost? Ti večno solnce, ti			jina je ležala tiha, ne	ema	
ConcDesc			- Admontan	and a state of the			
Visualize Original Conc.	Page 1 of 450 Go Next I	Last					
2							

www.clarin.si

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