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.

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

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 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 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)

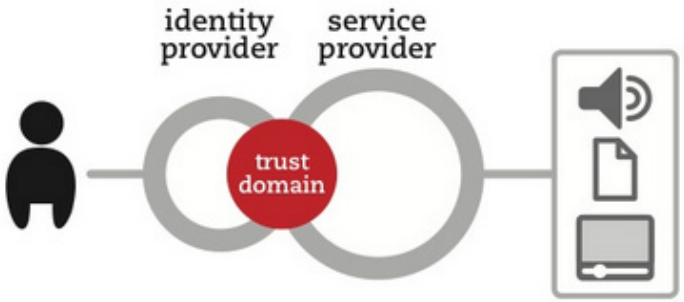
CLARIN.SI Cooperation

- 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

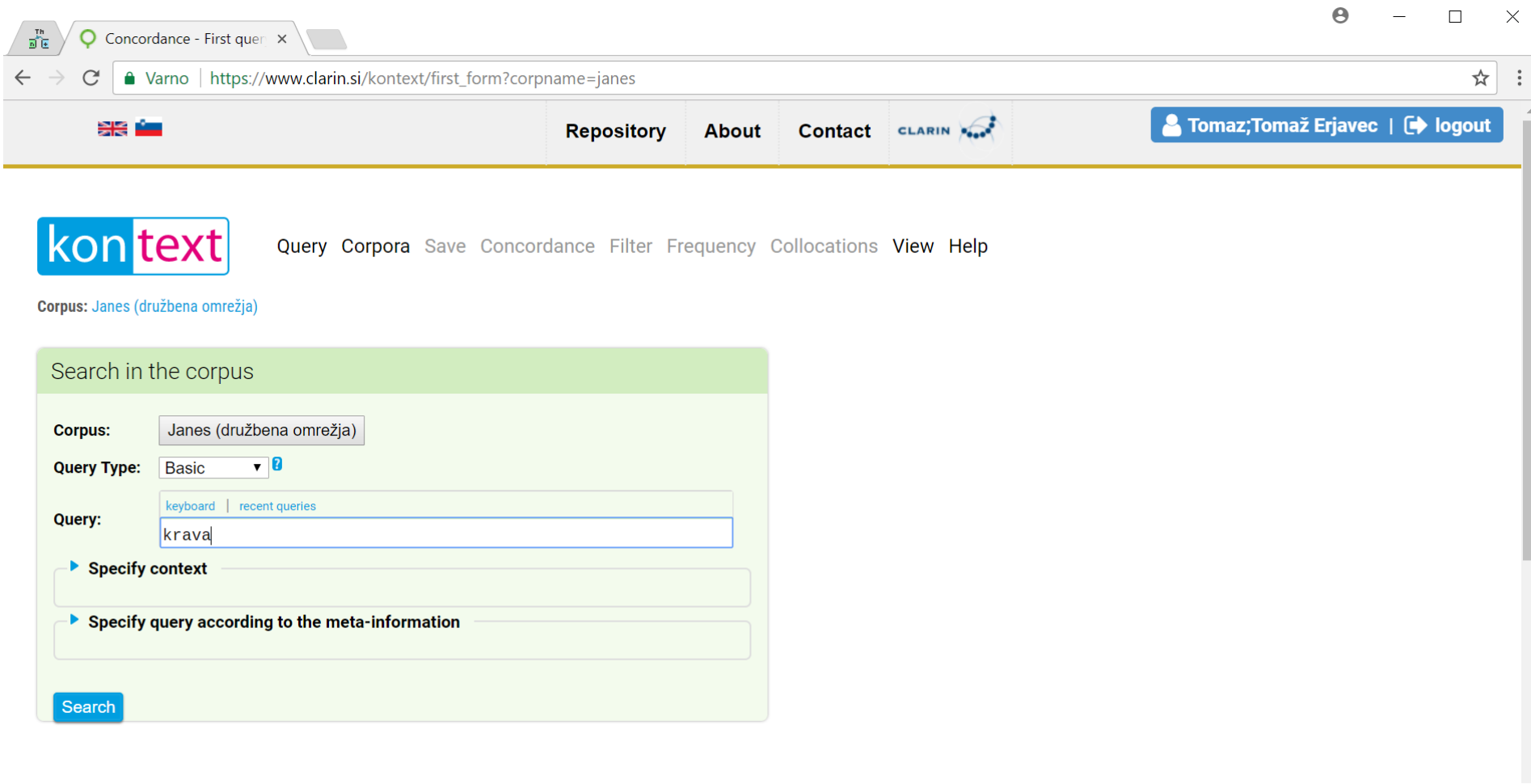


- Infrastructure for authentication and authorisation (AAI)
- Single Sign-On: separation between the Identity Provider and Service Provider
- As opposed to standard web login we here know the identity of the user
- Identity provider federations: EduGain
- Slovene users can, via EduGain, access most CLARIN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

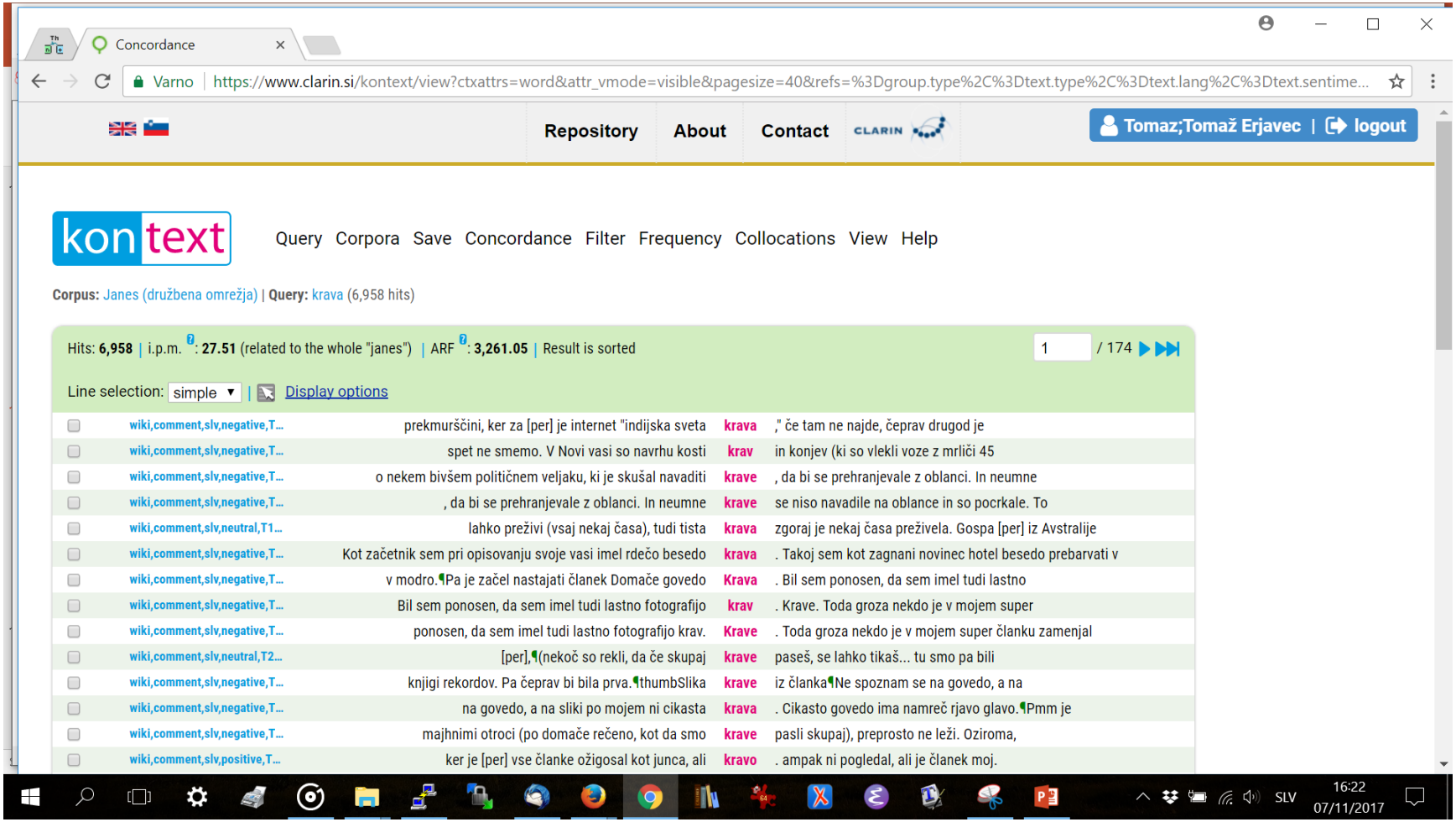
Concordancers

- 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)



Concordances



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 1 Go 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**

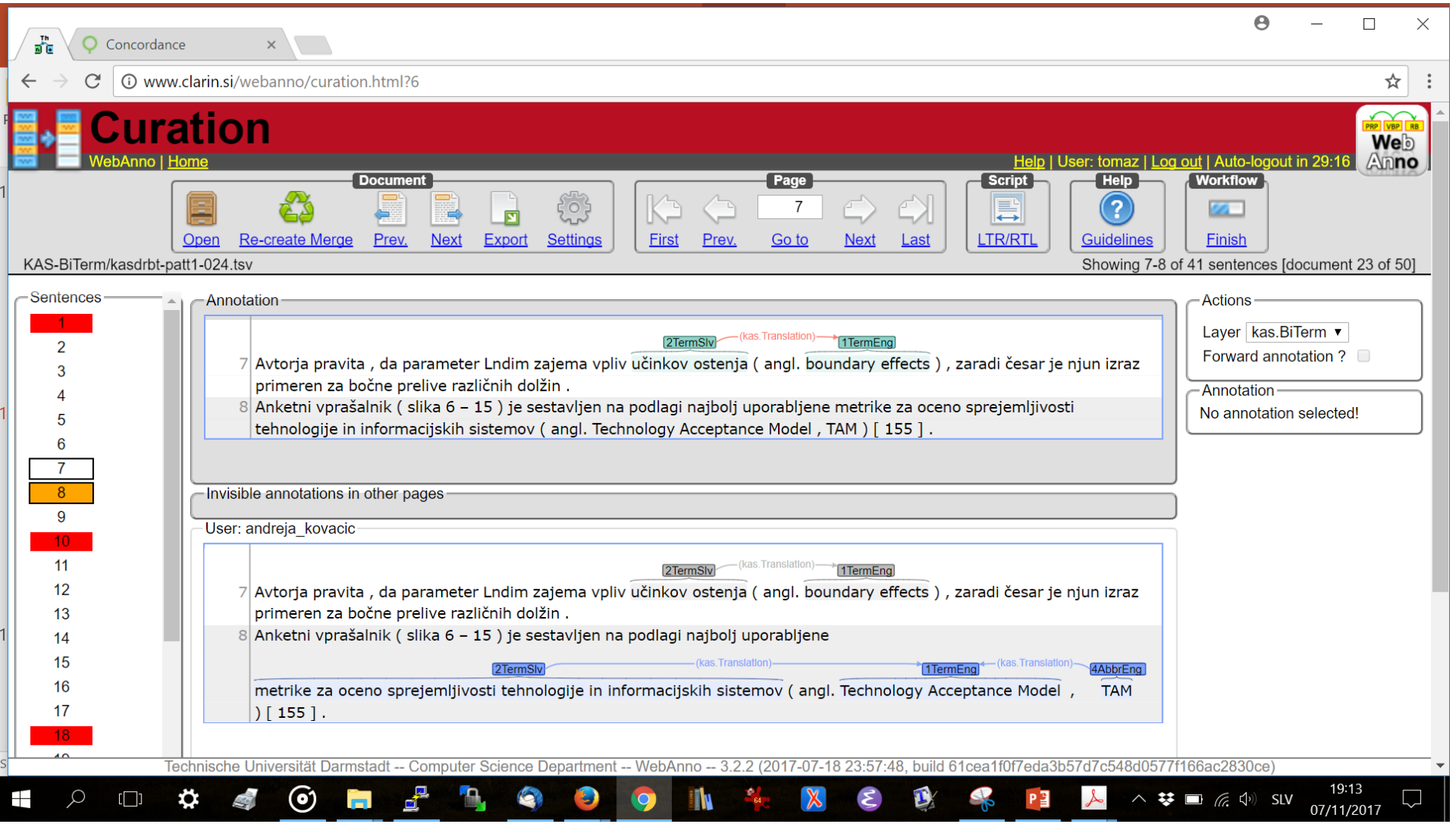
Reference corpus: **siParl 2.0 (parlament 1990-2018)**
Reference subcorpus: **Kučan**
Switch focus and reference (sub)corpus

Page 1 Go Next >

siParl 2.0 (parlament 1990-2018) : Janša

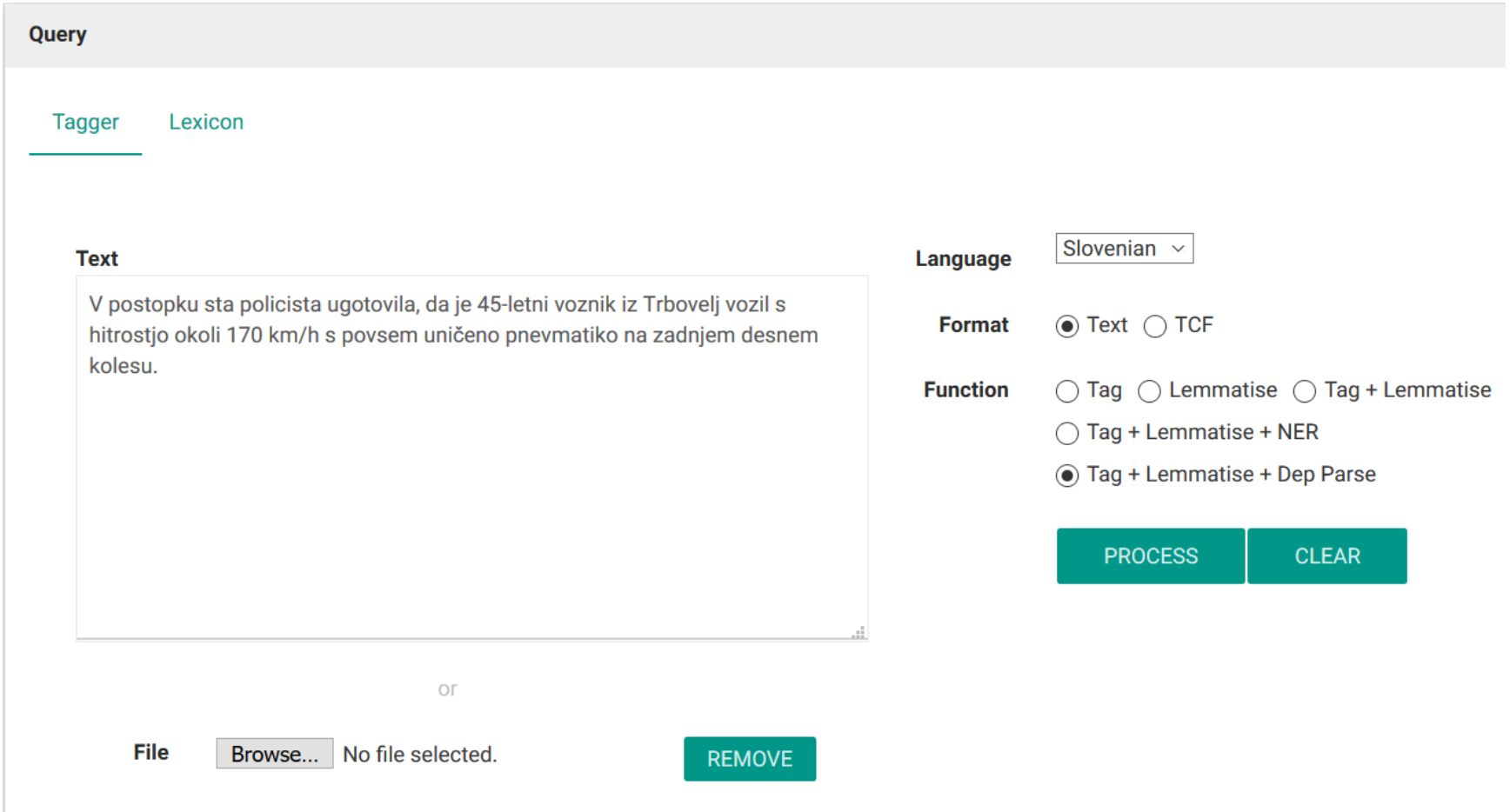
lemma	frequency	frequency/mill
nek	4,847	2752.3
narediti	1,812	1028.9
ukrep	1,396	792.7
delati	1,392	790.4
člen	1,188	674.6
točka	1,184	672.3
glasovati	1,019	578.6
tikati	1,010	573.5
odločba	980	556.5
mesec	962	546.3
stanje	921	523.0
ravno	872	495.1
kolega	872	495.1
situacija	838	475.8
predlagan	760	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)



ReLDI automatic text annotation

- Web form
- API



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