Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut „Jožef Stefan“

FRI, 10. 11. 2017

Pregled predavanja

1. Uvod
2. Raziskovalna infrastruktura CLARIN(.SI)
3. Storitve CLARIN.SI

Uvod

- Jezikovne tehnologije
  - glavna paradigma: nadzorovano strojno učenje
  - programi so večinoma jezikovno neodvisni
  - potrebujejo pa učne (ročno označene) jezikovne vire
  - + testne podatke
- Empirično podprte jezikoslovne raziskave:
  - temeljijo na realnih (in po možnosti označenih) besedilih
- Označene jezikovne vire potrebujemo za vsak jezik posebej
- Kje lahko dobimo take vire za slovenščino?

Jezikovni viri

1. Korpusi:
   - enovito kodirana in dokumentirana zbirka besedil
   - besedila (i)zbrana po vnaprej določenih kriterijih
   - označena
   - referenčni/specializirani; eno/večjezični; pisni/govorni
2. Leksikoni:
   - besedišče jezika
   - besede / fraze
   - oblikoslovje, skladnja, pomen, povezave, prevodi
3. Modeli:
   - podatki za nek program, ki mu omogoči označevanje besedil v nekem jeziku za neko raven označevanja
   - npr. CLOG model za lematizacijo slovenščine; Moses model za prevajanje slovenščina → angleščina

Ponovna uporaba

- Klasični pristop:
  - za vsako raziskavo posebej izdelati jezikovne vire
  - viri nedostopni drugim raziskovalcem
- Slabosti:
  - izdelava jezikovnega vira je lahko zelo draga in dolgotrajna, velika izguba časa in denarja, če se to počne večkrat
  - kasnejši raziskovalci ne morejo preveriti ali poboljšati prvih rezultatov
  - vzdržuje se monopol raziskovalcev oz. institucij, ki so vire izdelale
  - viri ne morejo biti uporabljeni pri razvoju produktov

Odprt dostop do rezultatov raziskovalnih projektov

- Brez ovir do publikacij in podatkov:
  - prihranek denarja in časa;
  - izogibanje ponavljanju dela;
  - spodbujanje sodelovanja;
  - večja transparentnost znanstvenega procesa;
  - spodbujanje inovacij
- Zelo močan trend v EU (H2020), tudi v Sloveniji
- Problemi pri omogočanju odprtega dostopa do jezikovnih virov:
  - avtorske pravice nad besedili
  - varovanje zasebnosti (tudi pravica do pozabe)
  - pogoji uporabe spletnih portalov (npr. Twitter)

Raziskovalne infrastrukture

Naprave, viri in storitve, ki jih uporablja znanstvena skupnost za izvajanje vrhunskih raziskav na svojih področjih.

Raziskovalne infrastrukture

- Pobuda EU: ESFRI (European Strategy Forum on Research Infrastructures) osnovan 2002
- Roadmap: predlagali 15 (2016: 21) RI, nekatere že delujejo kot ERIC (EU pravna oseba: European RI Consortium)
- Slovenija sodeluje v 14 RI (npr. CERN)
- Humanistika:
  - DARIAH ERIC / DARIAH-SI: Digitalna raziskovalna infrastruktura za umetnost in humanistiko (Digital Research Infrastructure for the Arts and Humanities)
  - **CLARIN ERIC / CLARIN.SI**: Infrastruktura za skupne jezikovne vire in tehnologije (Common Language Resources and Technology Infrastructure)

Common Language Resources and Technology Infrastructure

- Vizija: digitalni jezikovni viri in orodja za vse (evropske) jezike so dostopni prek enotne prijave za raziskovalce v humanistiki in družboslovju
- Namenjena dolgotrajnemu, obsežnemu in lahko dostopnemu hranjenju jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske dediščine
- Nova paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabnost in prilagajanja individualnim potrebam

CLARIN

Namen:

- obstoječa orodja in rešitve dati na voljo v enotni infrastrukturi
- omogočiti svetovalne in učne dejavnosti, kako orodja in vire prilagoditi specifičnim raziskovalnim potrebam
- prispevati k standardizaciji virov in orodij

CLARIN ERIC

- 19 držav članic + 2 opazovalki
- Sedež na Nizozemskem
- Podporno osebje, odbori za vodenje, delovne skupine
- Letne konference
- Virtual Language Observatory
- Večina dela se odvija v okviru nacionalnih konzorcijev

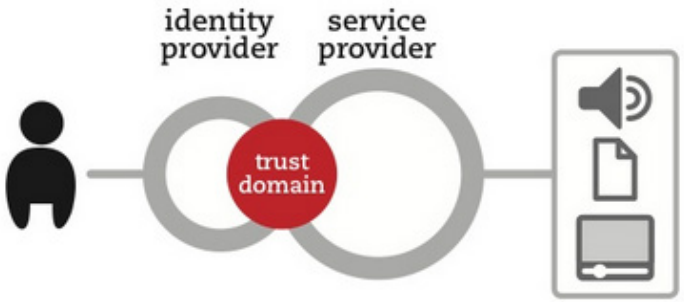
CLARIN.SI

- Začetek dela 2014
- Sedež IJS:
  - Odsek za tehnologije znanja
  - Laboratorij za umetni inteligenco
  - Center za mrežno infrastrukturo
- Organiziran kot konzorcij 12 parterjev:
  - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - 3 inštituti: ZRC SAZU, IJS, INZ
  - 3 društva: SDJT, DDR, Trojina
  - 2 podjetji: Amebis, Alpineon

CLARIN.SI

- Podpora dogodkom:
  - Seminar ReLDI “Empirični podatki v jezikoslovju: Od zasnove raziskave do analize podatkov” (FF, 21.–23. 6. 2017)
  - Redne konference „Jezikovne tehnologije in digitalna humanistika“ (…, 29. 9.–1. 10. 2016, 20.–21.9. 2018, …)
  - 4th Conference on CMC and Social Media Corpora for the Humanities (FF, 27.–28. 9. 2016)
- Spletne storitve:
  - konkordančniki (analize korpusov)
  - ročno označevanje (izdelava učnih korpusov)
  - delotoki (avtomatizirano označevanje)
- **Repozitorij**:
  - Dolgotrajno hranjenje jezikovnih virov in orodij

Prijava

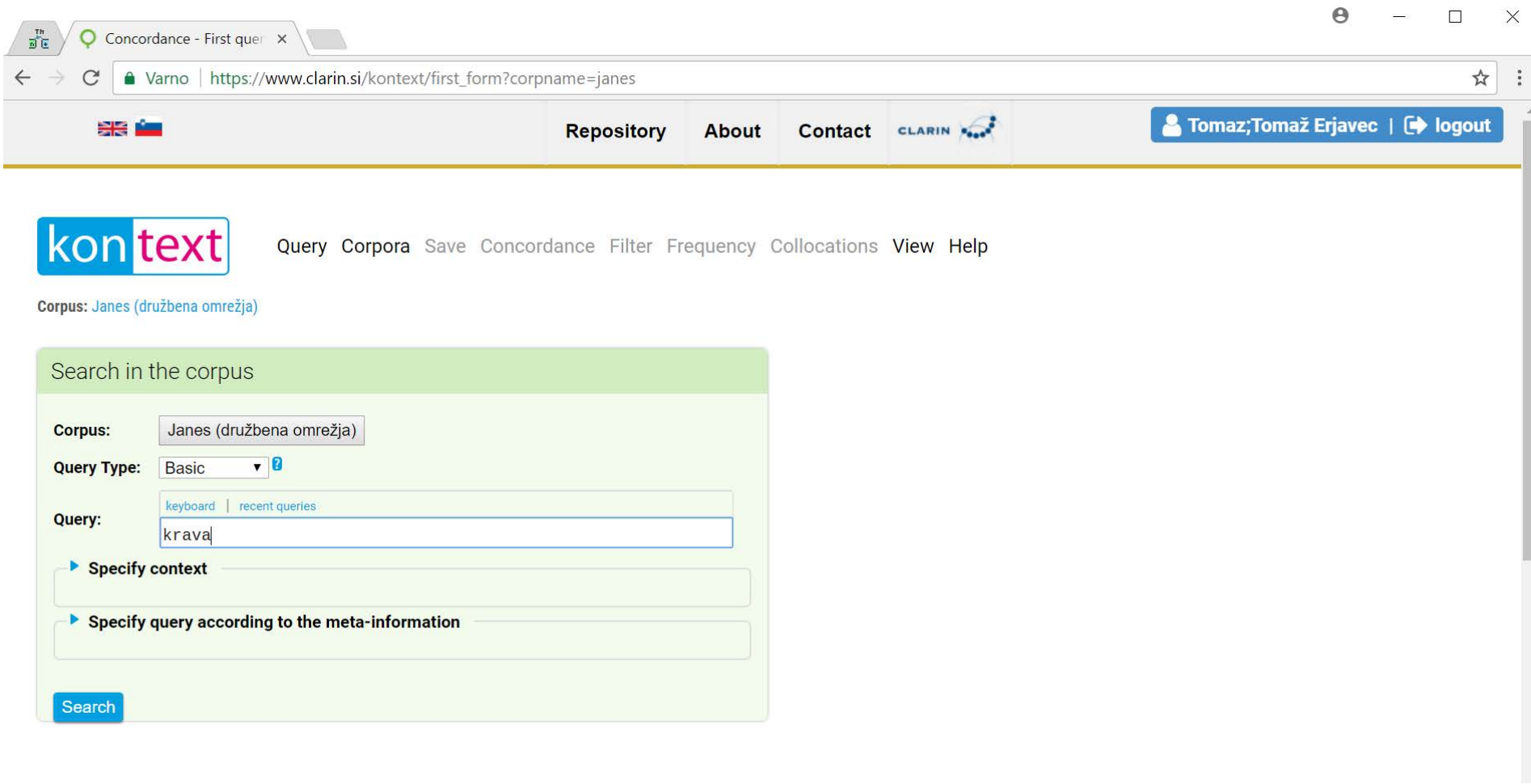


- Infrastruktura za avtentikacijo in avtorizacijo (AAI)
- Single Sign-On: ločevanje med ponudnikom storitve in ponudnikom identitete
- Za razliko od klasične spletne prijave je tu identiteta uporabnika poznana
- Federacije ponudnikov identitete
- EduGain: povezuje federacije ponudnikov identitete, da olajša dostop do vsebin, storitev in virov za globalno raziskovalno in izobraževalno skupnost
- Uporabite EduRoam uporabniško ime in geslo prek UL
- Slovenski uporabniki lahko dostopajo tudi do vseh servisov CLARIN po EU (npr. WebLicht)

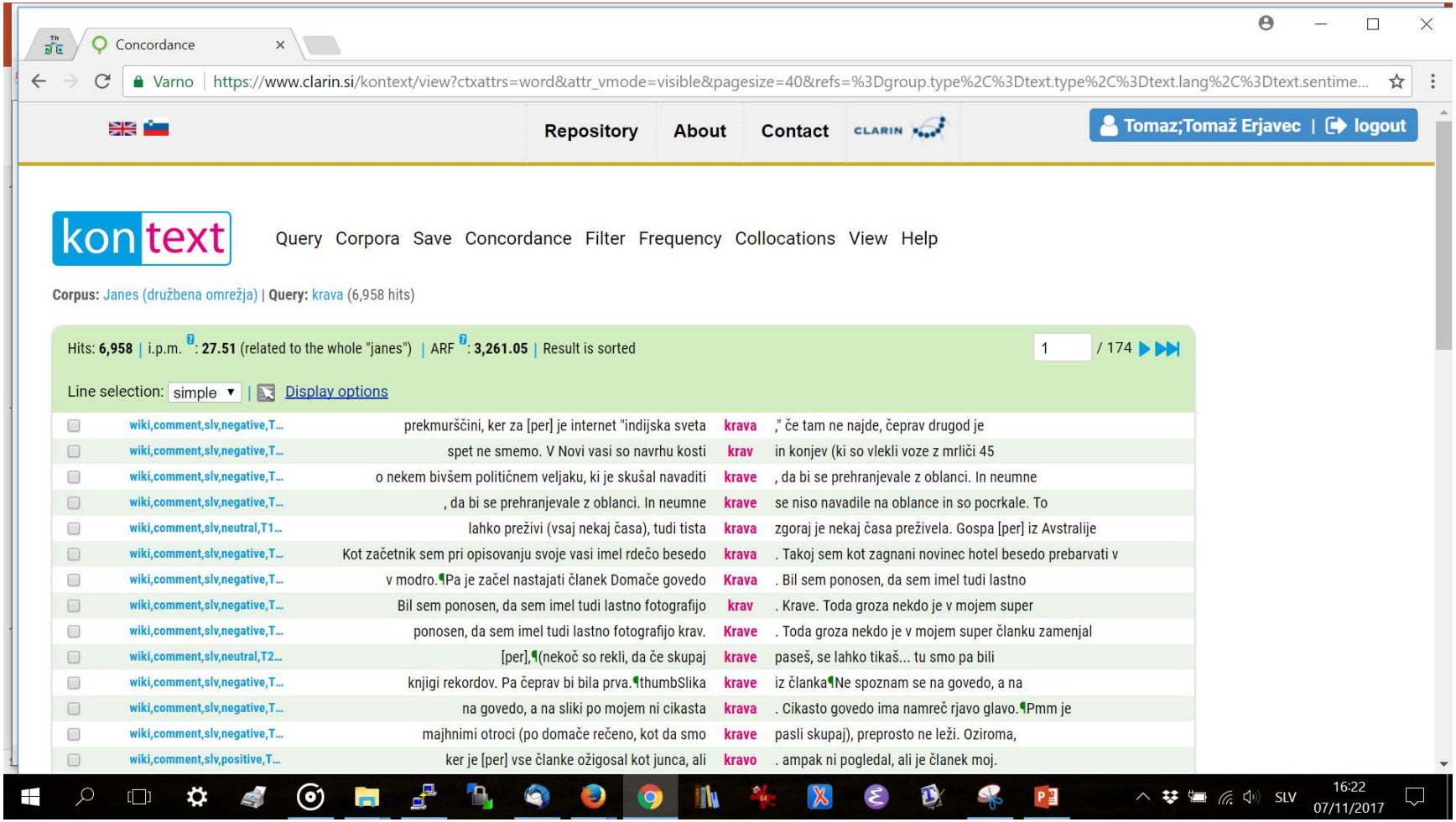
Konkordančniki

- KonText + noSketch Engine
- oba uporabljata isto zaledje: Manatee
- delujeta lahko z velikimi korpusi (več milijard besed)
- korpusi so lahko bogato označeni:
  - strukture in metapodatki (datum objave, vrsta besedila, spol avtorja, standardnost besedila, …)
  - atributi besed (oblikoskladenjska oznaka, lema, normalizirana oblika, …)
- bogat poizvedovalni jezik CQL
- nastavitve pogleda
- raznovrstni izpisi in analize
- CLARIN.SI trenutno ponuja okoli 50 korpusov

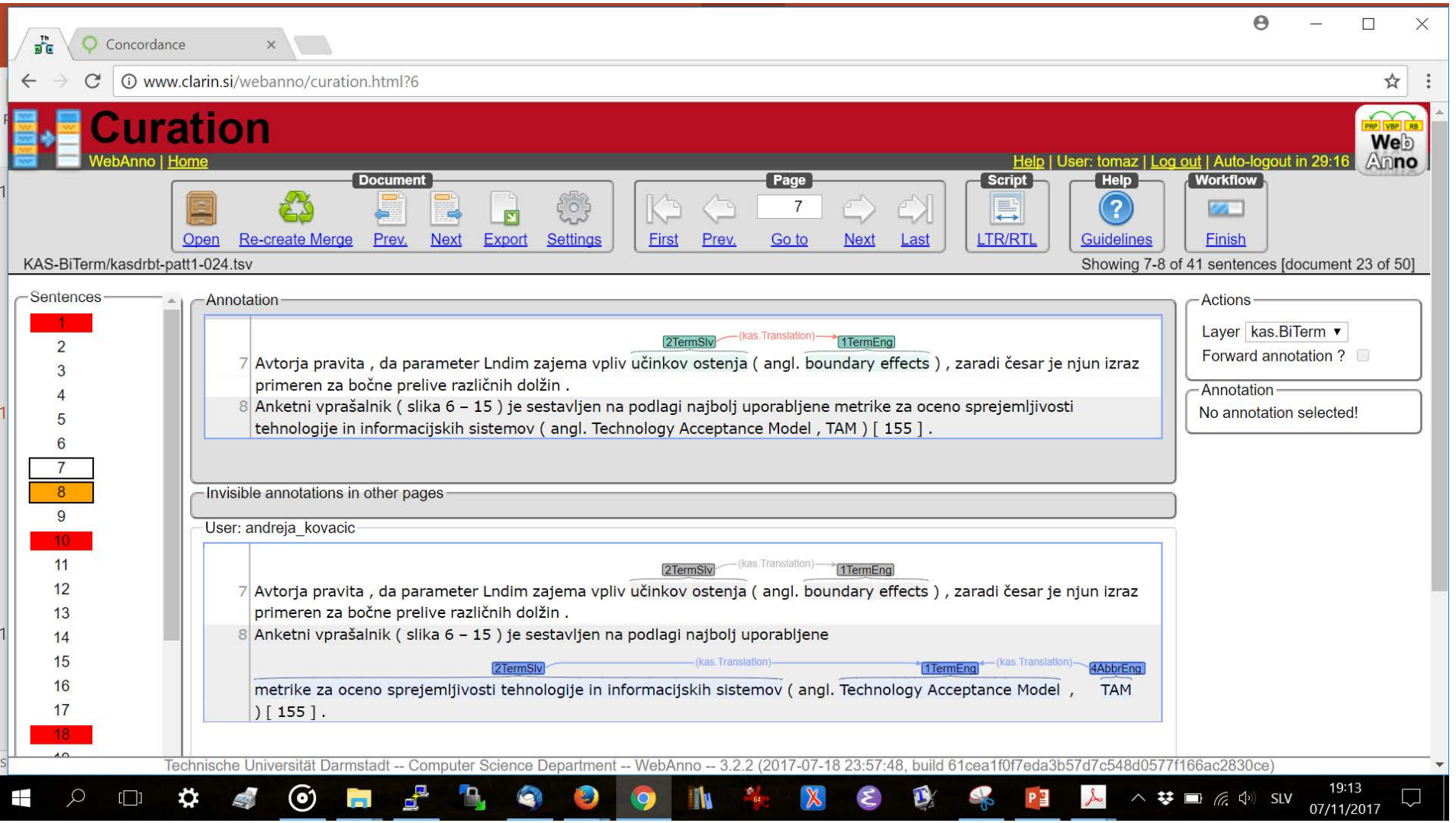
KonText (CLARIN-CZ)



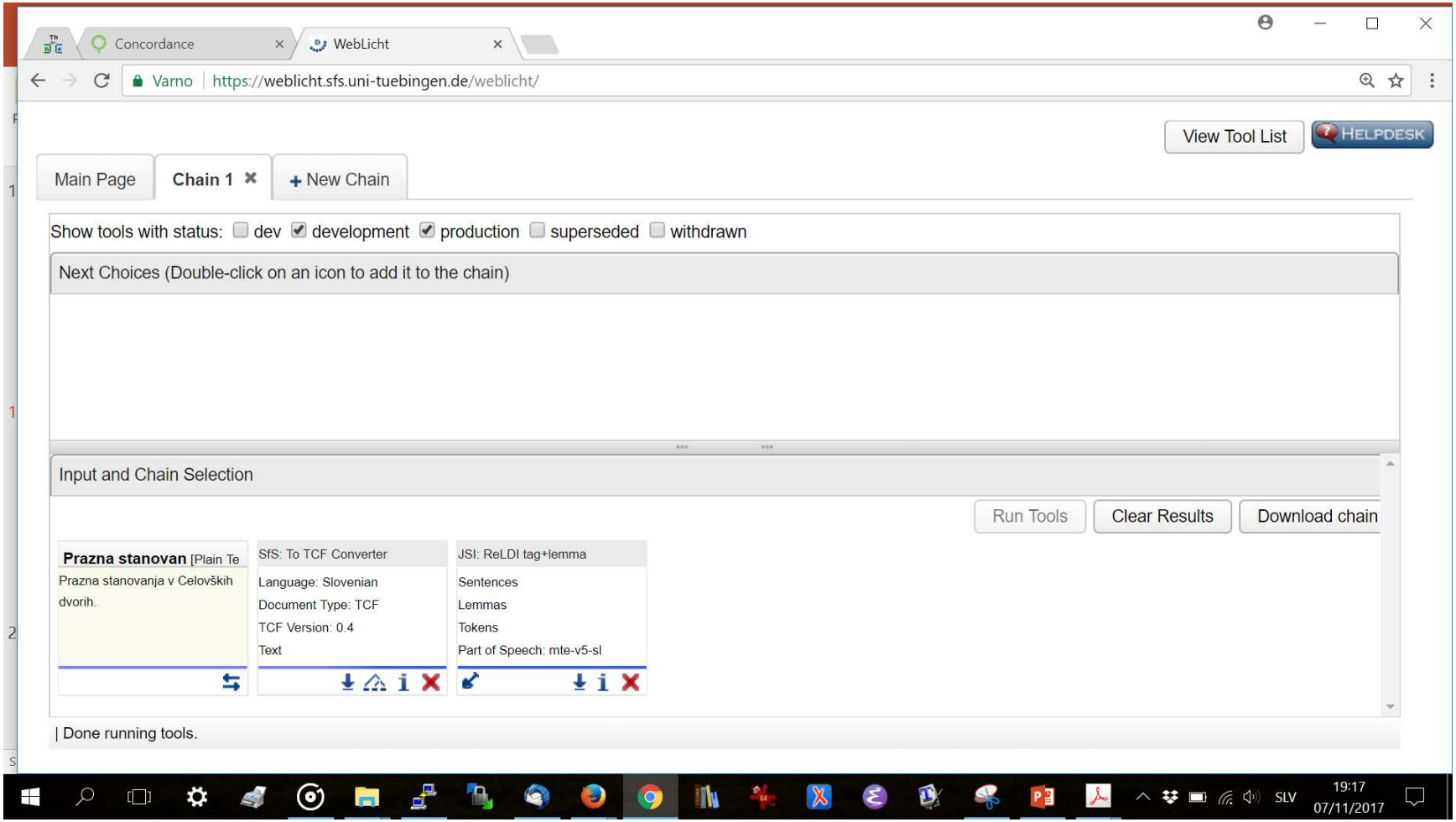
Konkordance



WebAnno (CLARIN-DE)



WebLicht (CLARIN-DE)



Nacionalni repozitoriji

- Zaenkrat najbolj pomembna storitev CLARIN-a
- Stalna in varna hramba jezikovnih virov (https, Nagios)
- Eksplicitni pogoji uporabe (ToS, licenca)
- Etični kodeks (CoCo)
- Standarden zapis metapodatkov:
  - Component Metadata Infrastructure (CMDI)
  - Dublin Core (DC)
- Izvoz metapodatkov (metadata harvesting)

Repozitorij CLARIN.SI

- Osnovan na platformi DSpace, namenjena odprtim digitalnim repozitorijem
- DSpace prilagojen za namene CLARIN repozitorijev v češkem CLARIN (t.i. DSpace/LINDAT)
- Poleg Slovenije ga uporablja tudi Češka, Norveška, Poljska, Italija, …
- Vzdrževanje Dspace/LINDAT na GitHub, veja za CLARIN.SI na GitLab@IJS

Certifikati

CLARIN.SI

2014 DSA 2017

CLARIN CENTRE B

- Ceritificiran B center CLARINa
- Certifikat DSA
- Ponudnik metapodatkov prek OAI-MHP
- Zaveden v registru re3data
- Povezan v sledenje rezultatom EU projektom OpenAIRE

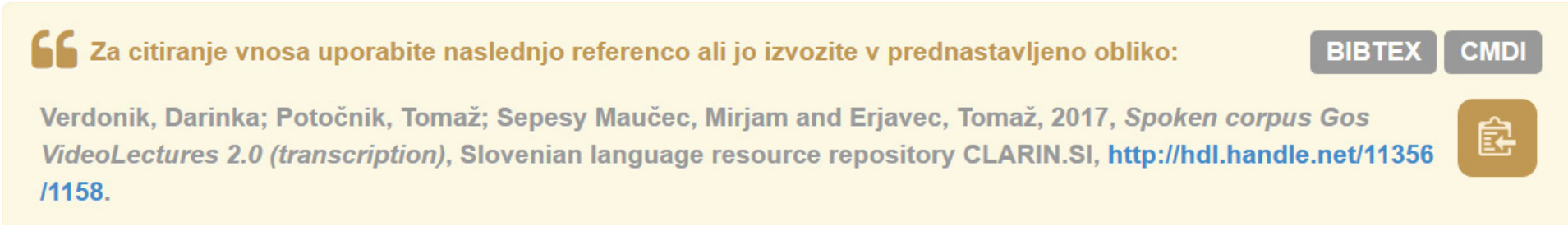
OPEN ARCHIVES

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

OpenAIRE

Stalni identifikatorji in citiranje virov

- Problem stabilnih identifikatorjev / URLjev, ki jih je možno dolgoročno citirati
- Najbolj razširjena rešitev: DOI
- CLARIN uporablja sistem Handle
- **http://hdl.handle.net/11356/1044** → https://www.clarin.si/repository/xmlui/handle/11356/1044



Zaključki

- Namen CLARIN(.SI) je spodbujati raziskave, ki potrebujejo dostop do jezikovnih podatkov
  - digitalna humanistika in družboslovje
  - jezikovne tehnologije
  - vse ostale vede, kjer je jezik pomemben
- Odprt dostop do virov, orodij in storitev
- Kjer je potrebna avtentikacija, se uporablja AAI
- Stabilno citiranje skozi sistem Handle