Konkordančniki CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Predavanje na doktorskem študiju ZRC SAZU
2022-05-23

Pregled predavanja

Uvod: CLARIN(.SI)

CLARIN: Common Language Resources and Technology Infrastructure

- CLARIN ERIC: Evropska raziskovalna infrastruktura
- 20 držav članic, mdr. Slovenija
- Vizija: digitalni jezikovni viri in orodja za vse (evropske) jezike so dostopni za raziskovalce v humanistiki in družboslovju
- Namenjena je dolgotrajnemu in obsežnemu hranjenju ter dostopu do jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske kulturne dediščine
- Nova paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabnosti in prilagajanja individualnim potrebam

CLARIN.SI

`http://www.clarin.si/`

- Začetek dela v 2014-2015
- Institut "Jožef Stefan" + konzorcij 12 partnerjev
- Trije stebri delovanja CLARIN.SI:
  1. Repozitorij jezikovnih virov (in orodij)
  2. **Dva (oz. trije) konkordančniki** in druge spletne storitve
  3. Podpora raziskovalnim dejavnostim (vsebinska in finančna)

Trije konkordančniki CLARIN.SI

Konkordančniki v Sloveniji

- Najbolj znan je konkorančnik Gigafide
  in drugi konkordančniki oz. korpusi CJVT: Gos, Šolar, Lektor
- Sektor za prevajanje Generalnega sekretariata Vlade RS:
  Evroterm, kombinacija terminološke baze in vzporednega korpusa
- ISJFR ZRC: Nova Beseda

Večinoma: 1 konkordančnik za 1 korpus

Corpus Workbench (CWB)

- Christ, Oliver (1994). A modular and flexible architecture for an integrated corpus query system. Papers in Computational Lexicography (COMPLEX '94), Budimpešta.
- Vzdrževanje in razvoj CWB prevzel Stefan Evert, kasneje skupaj z Andrew Hardiejem
- Splošen sistem za poljubne korpuse
- Po svetu več instalacij CWB, večinoma za interne uporabnike
- CWB =
  - zaledni del CQP: hiter dostop do rezultatov poizvedbe
  - čelni del CQPweb: spletni vmesnik, ki ga vidi uporabnik
- CQP bil dolog let instaliran tudi na nl.ijs.si, razvili smo svoj čelni del CUWI

Vertikalne datoteke

```
<text corpus_id="imp" corpus="IMP (starejša besedila)"
      info="https://www.clarin.si/noske/...corpname=imp"
      id="ZRC_00001-1584" title="Biblija (vzorec)"
      author="Dalmatin, Jurij" year="1584">
<gap/>
<p id="ZRC_00001-1584.p2">
<s>
INu         in          in          Cc        Vp
on          on          on          Pp3msn    Zotmei
je          je          biti        Va-r3s-n  Gp-ste-n
ſvoje       svoje       svoj        Px-nsa    Zp-set
dvanajſt    dvanajst    dvanajst    Mlc-pa    Kbg-mt
Iogre       jogre       joger       Ncmpa     Sommt
k'          k           k           Sd        Dd
ſebi        sebi        se          Px---d    Zp---d
poklizal    poklical    poklicati   Vmep-sm   Ggdd-em
<g/>
,           ,           ,           Z         U
```

- Konfiguracijska datoteka korpusa definira strukturne oznake, njihove atribute ter imena pozicijskih oznak
- Korpus ima tako lahko poljubne oznake

Corpus Query Language (CQL)

Bogat poizvedovalni jezik:

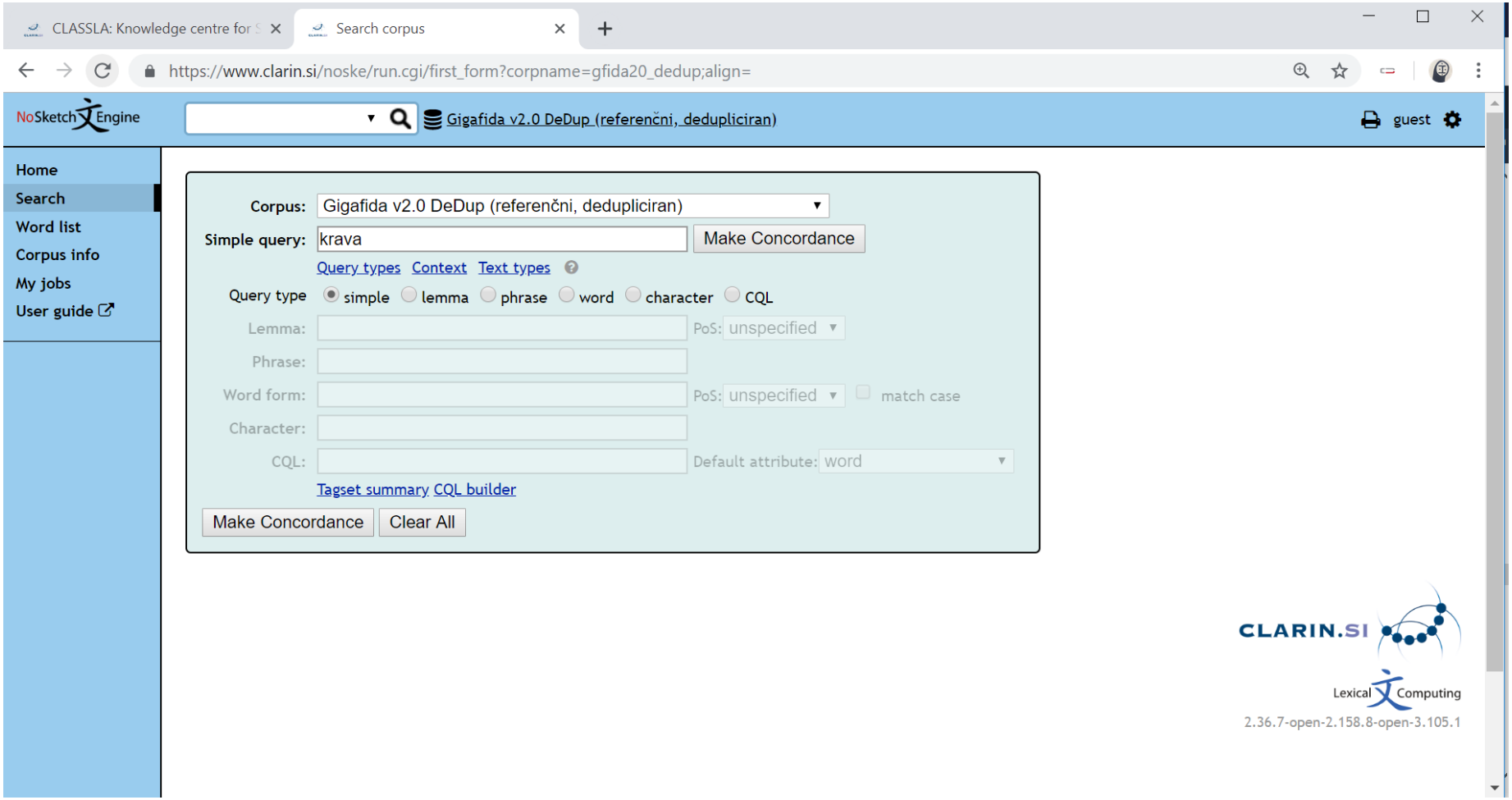
- Iskanje po poljubni pozicijski oznaki, npr. `[lemma="krava"]`
- Regularni izrazi, npr. `[lemma="krava.*"]`
- Logične kombinacije pogojev, npr.
  `[lemma="krava.*" & word!="krava.*"]`
- Iskanje več pojavnic, npr.
  `[lemma="zelo"] []{0,2} [lemma="krava"]`
- Omejitve glede na strukturne oznake
  `[lemma="krava"] within <text year="2011"/>`

Večino vrst iskanj "na klik" z vmesnika je mogoče prevesti v CQL

Sketch Engine in noSketch Engine

- Lexical Computing (Brno): podjetje, ki ponuja komercialen konkordančnik Sketch Engine (SkE)
- Za SkE reimplementacija CQP zalednega (Manatee) in čelnega CQPweb (Bonito) dela:
  Rychlỳ, Pavel. Manatee/Bonito-A Modular Corpus Manager. In: Recent Advances in Language Processing (RASLAN 2007).
- Manatee + Bonito = osnova komercialnega konkordančnika Sketch Engine
- Format vhodnih podatkov in poizvedovalni jezik sta ostala nespremenjena glede na CWB
- Na voljo odprtodostopna različica Sketch Engine, imenovana noSketch Engine
- noSkE ne vsebuje nekaterih naprednih funkcionalnosti SkE
- noSkE ne podpora prijave uporabnikov

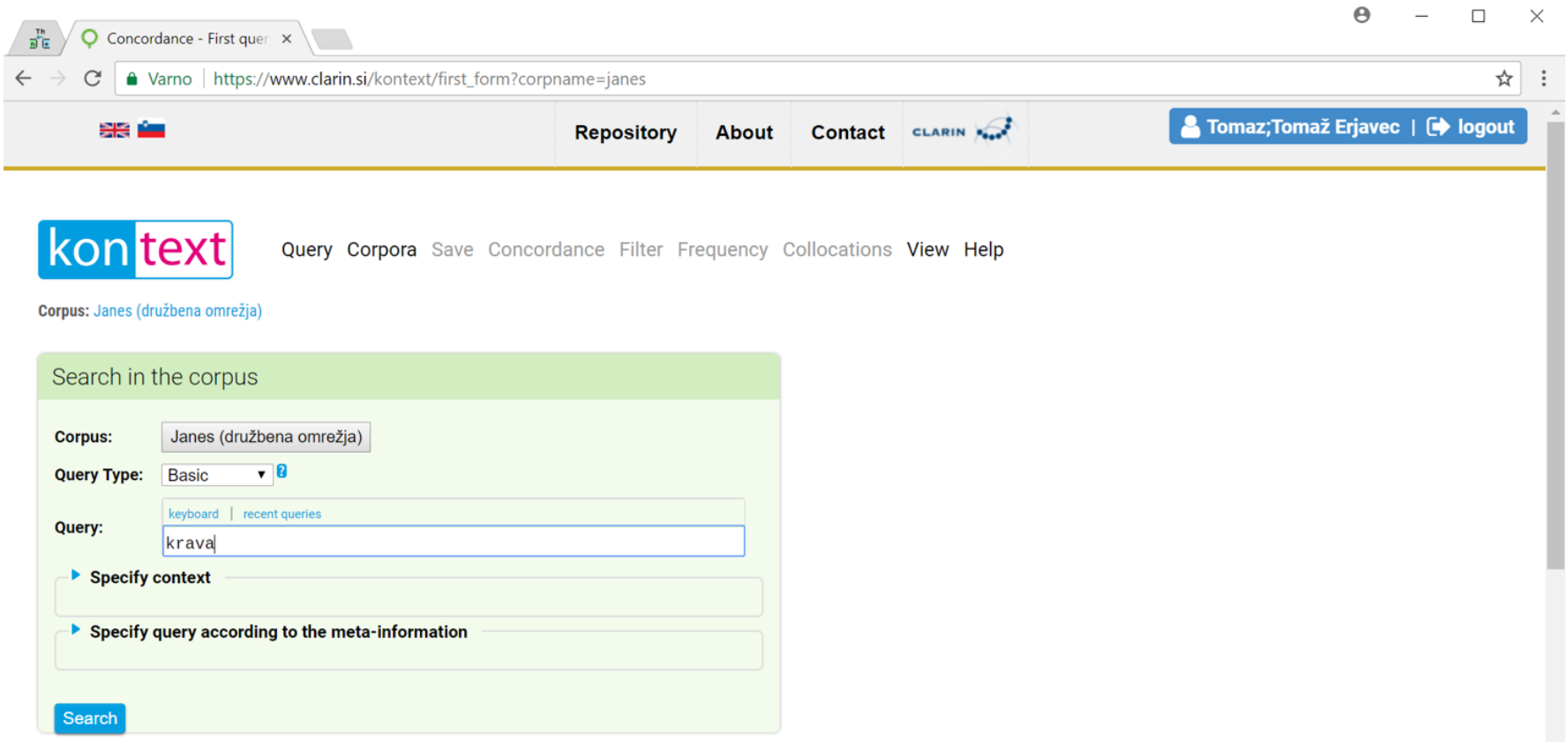
noSketch Engine Bonito



KonText

- KonText je konkordančnik, ki so ga na Karlovi univerzi razvili za Češki nacionalni korpus
- Za zaledni del uporablja Manatee, čelni del napisan na novo
- Drugačna razporeditev menijev in funkcij
- Glavna prednost: podpira prijavo v sistem
  - prijava AAI (EduGain / EduRoam)
  - shranjene poizvedbe, nastavitve zaslona, podkorpusi

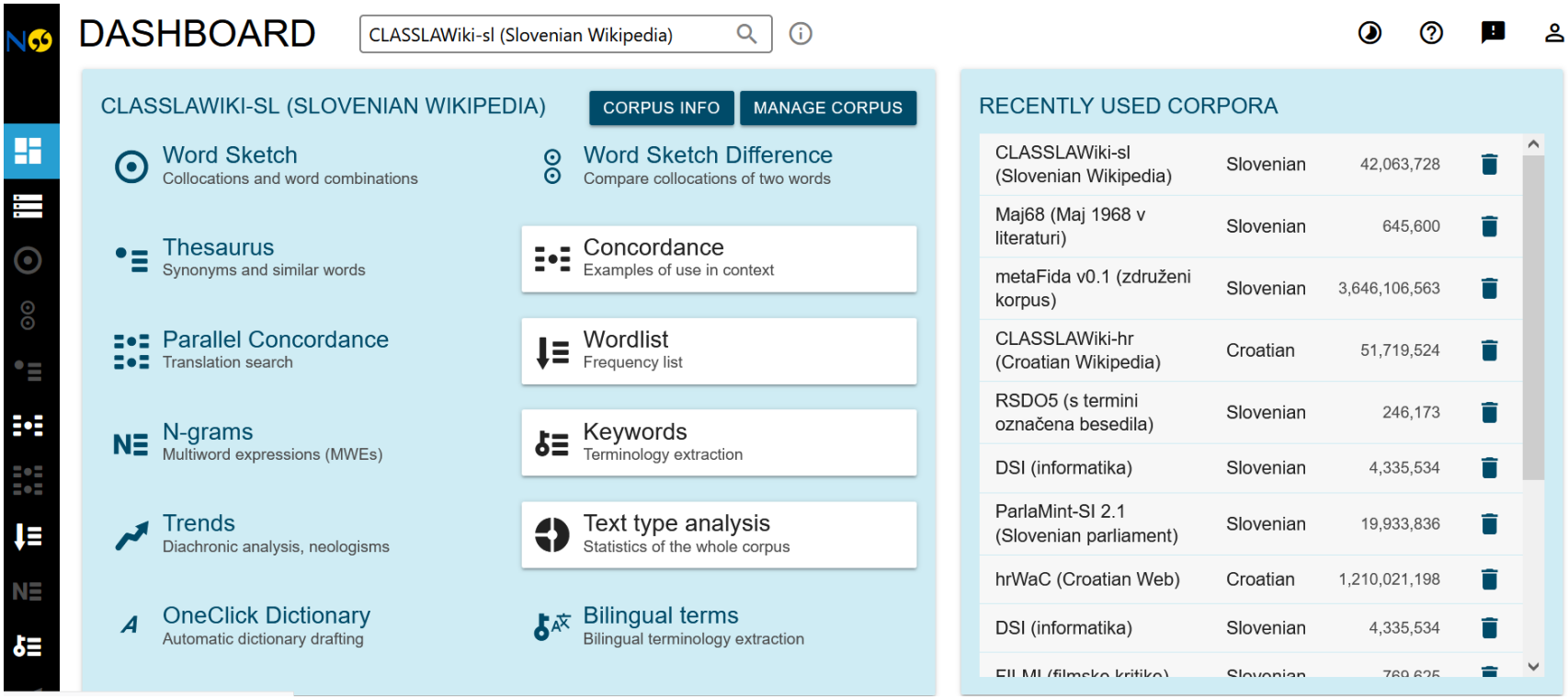
KonText


Concordance - First quer
Varno
https://www.clarin.si/kontext/first_form?corpname=janes
Repository
About
Contact
CLARIN
Tomaz;Tomaž Erjavec | logout
kontext
Query Corpora Save Concordance Filter Frequency Collocations View Help
Corpus: Janes (družbena omrežja)
Search in the corpus
Corpus:
Janes (družbena omrežja)
Query Type:
Basic
Query:
keyboard | recent queries
krava
Specify context
Specify query according to the meta-information
Search

Novi noSketch Engine (Crystal)

- Par let nazaj je Lexical Computing radikalno spremenil čelni del Sketch Engine:
  - bistveno drugačen izgled
  - drugačna razporeditev menijev in funkcij
- Nova verzija Manatee in Bonito
- Dodatna plast na čelnem delu: Crystal
- Stari Manatee in Bonito nista več vzdrževana

noSketch Engine Crystal



- CLARIN.SI Crystal še ni bil objavljen
- `https://www.clarin.si/ske-beta/`
- Uporabniško ime: dev
- Geslo: alfabetagama

Primerjava

	Bonito	Crystal	KonText
Dokumentacija	Ne	Da	Ne
Slovenski vmesnik	Da	Ne	Da
Ključne besede	Da	Da	Ne
Vzdrževan	Ne	Da	Da
Mogoča prijava	Ne	Ne	Da
SkE kompatibilnost	Ne	Da	Da
CNC kompatibilnost	Ne	Ne	Da
Izvoz XML	Da	Ne	Ne
Povezave z viri IJS / CLARIN.SI	Da	Ne	Ne

Kratek pregled korpusov

Korpusi

- Konkordančniki omogočajo dostop do cca. 100 korpusov v 33 jezikih z 20 milijardami besed
- Korpusi so zelo raznovrstni, npr.
  - Referenčni: GigaFida
  - Znanstvena besedila: KAS
  - Govorni: Gos, GosVL
  - Starejša slovenščina: IMP
  - Uporabniško generirane vsebine: Janes
  - Vzporedni: EU-DGT
  - Drugi južnoslovanski jeziki: hrvaški, srbski, makedonski, črnogorski
  - Drugi jeziki: angleščina, japonščina

Oznake korpusov

- Korpusi se med seboj zelo razlikujejo, zato se razlikujejo tudi oznake
- Nekaj truda, da so si vseeno čim bolj podobne:
  - `<text>`, `<p>`, `<s>`
  - `word`, (`norm`), `lemma`, `tag`, `tag-en`
- Vendar so atributi besedil zelo raznovrstni
- Nekateri korpusi imajo še nadaljnje oznake (Universal Dependencies)

MetaFida

- MetaFida: združeneni korpus slovenskih korpusov na konkordančnikih (projekt RSDO)
- Izbira korpusov (=34 korpusov, 3.5 milijarde besed):
  - Čim večje število čim bolj raznovrstnih korpusov
  - Vključeni taki, ki niso podmnožica drugih korpusov
  - Pojavnice označene vsaj z lemo in oblikoskladenjsko oznako
- Odstranjevanje podvojenih odstavkov: zbrisanih 12,5 % odstavkov in 6,5 % besedil
- Sortirano po letnici besedila (tisti brez letnice na koncu)
- = različica 0.1 korpusa MetaFida
- Različica 1.0 ob koncu projekta RSDO

Zaključki

Zaključki

- Predstavil zgodovino in trenutno stanje konkordančnikov, ki jih ponuja CLARIN.SI
- in kratek pregled dostopnih korpusov
- Dosti korpusov, ki so dostopni na konkordančnikih je dostopnih tudi za prevzem v repozitoriju CLARIN.SI
- Nadaljnje delo:
  - Dokončati instalacijo Crystal in ga objaviti
  - Lokalizacija Crystal v slovenščino
  - Novi korpusi, predvsem projekta RSDO
  - Tečaj(i) uporabe konkordančnikov

Konkordančniki CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Predavanje na doktorskem študiju ZRC SAZU
2022-05-23