

Uporaba konkordančnika noSketch Engine @ CLARIN.SI

Tomaž Erjavec

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

Delavnica

"Uporaba digitalizirane zbirke pregovorov Pregovori 1.0 na CLARIN.SI"
2023-03-15
ZRC SAZU

Pregled predavanja

- 1 Uvod: CLARIN(.SI)
- 2 Trije konkordančniki CLARIN.SI
- 3 Vertikalne datoteke in CQL
- 4 Zaključki

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



<https://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 **Trije konkordančniki** in druge spletne storitve
 - 3 Podpora raziskovalnim dejavnostim (vsebinska in finančna)

Trije konkordančniki CLARIN.SI

Sketch Engine in noSketch Engine

- Lexical Computing (Brno): podjetje, ki ponuja komercialni konkordančnik Sketch Engine (SkE)
- SkE je sestavljen iz zaledenega (Manatee) in čelnega dela (Bonito)
- Na voljo odprtodostopna različica Sketch Engine, imenovana noSketch Engine (noSkE)
ne vsebuje nekaterih naprednih funkcionalnosti SkE
- Pred 5 leti je Lexical Computing bistveno spremenil čelni del: Crystal
 - bistveno drugačen izgled
 - drugačna razporeditev menijev in funkcij
- 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

noSketch Engine Bonito

CLASSLA: Knowledge centre for ... Search corpus

https://www.clarin.si/noske/run.cgi/first_form?corpname=gfida20_dedup;align=

NoSketch Engine

Gigafida v2.0 DeDup (referenčni, dedupliciran) guest

Home
Search
Word list
Corpus Info
My jobs
User guide

Corpus: Gigafida v2.0 DeDup (referenčni, dedupliciran)

Simple query: krava Make Concordance

Query types Context Text types

Query type: ☒ simple ☐ lemma ☐ phrase ☐ word ☐ character ☐ CQL

Lemma: PoS: unspecified

Phrase: PoS: unspecified match case

Word form: PoS: unspecified match case

Character: Default attribute: word

CQL: Default attribute: word

Tagset summary CQL builder

Make Concordance Clear All

CLARIN.SI
Lexical Computing
2.36.7-open-2.158.8-open-3.105.1

KonText

The screenshot shows a web browser window with the KonText application. The browser's address bar shows the URL `https://www.clarin.si/kontext/first_form?corpname=janes`. The application's header includes a navigation bar with links for Repository, About, and Contact, and a user profile section for Tomaz Tomaž Erjavec with a logout button. The main content area features the KonText logo and a navigation menu with links for Query, Corpora, Save, Concordance, Filter, Frequency, Collocations, View, and Help. Below the navigation menu, the current corpus is identified as 'Janes (družbena omrežja)'. A search form is displayed with the following fields: Corpus (Janes (družbena omrežja)), Query Type (Basic), and Query (krava). The search form also includes options for keyboard and recent queries, and buttons for specifying context and meta-information. A Search button is located at the bottom of the search form.

Concordance - First query x

Varno | https://www.clarin.si/kontext/first_form?corpname=janes

Repository About Contact CLARIN

Tomaz; Tomaž Erjavec | logout

kon text Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: [Janes \(družbena omrežja\)](#)

Search in the corpus

Corpus:

Query Type: ⓘ

keyboard recent queries

Query:

Specify context

Specify query according to the meta-information

Search

noSketch Engine Crystal



DASHBOARD

CLASSLAWiki-sl (Slovenian Wikipedia) 🔍 ⓘ

CLASSLAWIKI-SL (SLOVENIAN WIKIPEDIA)

CORPUS INFO

MANAGE CORPUS

**Word Sketch**
Collocations and word combinations

**Thesaurus**
Synonyms and similar words

**Parallel Concordance**
Translation search

**N-grams**
Multiword expressions (MWEs)

**Trends**
Diachronic analysis, neologisms

**OneClick Dictionary**
Automatic dictionary drafting

**Word Sketch Difference**
Compare collocations of two words

**Concordance**
Examples of use in context

**Wordlist**
Frequency list

**Keywords**
Terminology extraction

**Text type analysis**
Statistics of the whole corpus

**Bilingual terms**
Bilingual terminology extraction

RECENTLY USED CORPORA

CLASSLAWiki-sl (Slovenian Wikipedia)	Slovenian	42,063,728	
Maj68 (Maj 1968 v literaturi)	Slovenian	645,600	
metaFida v0.1 (združeni korpus)	Slovenian	3,646,106,563	
CLASSLAWiki-hr (Croatian Wikipedia)	Croatian	51,719,524	
RSDO5 (s termini označena besedila)	Slovenian	246,173	
DSI (informatika)	Slovenian	4,335,534	
ParlaMint-SI 2.1 (Slovenian parliament)	Slovenian	19,933,836	
hrWaC (Croatian Web)	Croatian	1,210,021,198	
DSI (informatika)	Slovenian	4,335,534	
CIJ MI (Ekonomska kritika)	Slovenian	760,626	



Primerjava

	Bonito	Crystal	KonText
Dokumentacija	Ne	Da	(Da)
Slovenski vmesnik	Da	Ne/(Da)	Da
Ključne besede	Da	Da	Ne
Prikaz govora	(Ne)	(Ne)	Da
Vzdrževan	Ne	Da	Da
Prijava	Ne	Ne/(Da)	Da
Podkorpusi	Da	Ne/(Da)	Da
Osebnе nastavitve	Da	Ne/(Da)	Da
API XML	Da	Ne	Ne
API JSON	Da	Da	Ne
Povezave	Da	Ne	Ne

Korpusi

- 100 korpusov, 33 jezikov, "23" milijard besed
 - Referenčni: GigaFida, metaFida (RSDO)
 - Znanstvena besedila: KAS / OSS (RSDO)
 - Govorni: Gos 2.0, /Artur 1.1/ (RSDO)
 - Starejša slovenščina: IMP (Balkan-Kommission in Google), ELTeC-slv (COST)
 - Uporabniško generirane vsebine: Janes (ARRS)
 - Ostali južnoslovanski jeziki (CLARIN.SI CLASSLA K-centre): hbs, makedonski, bolgarski
 - Drugi jeziki: angleščina, japonščina
 - Vzporedni: EU-DGT, Trans5
- Majhni, a (delno) ročno označeni korpusi: ssj500k / SUK (RSDO); Janes-Tag/Norm 3.0 (RSDO)
- Večji avtomatsko označeni: točnost in priklic oznak nista 100 %!

Vertikalne datoteke in CQL

Vertikalne datoteke

- Konfiguracijska datoteka korpusa definira strukturne oznake, njihove attribute ter imena pozicijskih oznak
- Imena struktur, atributov in pozicij so poljubna
- Incipit vertikalne datoteke korpusa MetaFida 0.1:

```
<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
fvoje    svoje    svoj    Px-nsa   Zp-set
dvanajft dvanajst dvanajst Mlc-pa   Kbg-mt
Iogre    jogre    joger   Ncmpa    Sommt
k'       k        k        Sd       Dd
febi     sebi     se       Px---d   Zp---d
poklizar poklical poklicati Vmep-sm  Ggdd-em
<g/>
,        ,        ,        Z        U
```

Corpus Query Language (CQL)

Bogat poizvedovalni jezik:

- Iskanje po poljubni pozicijski oznaki: `[lemma="krava"]`
- Regularni izrazi: `[lemma="krav.*"]`
- Logične kombinacije pogojev:
`[lemma="krava" & word!="krava"]`
- Iskanje več pojavnic:
`[lemma="zelo" [] {0,2} [lemma="krava"]`
- Omejitve glede na strukturne oznake:
`[lemma="krava"] within <text year="2011"/>`
- Prikaz vsebine strukturnih oznak:
`<name type="ORG"/> containing [lemma="Slovenija"]`

Večino vrst iskanj "na klik" je mogoče prevesti v ukaz CQL

Oznake korpusov

- Korpusi se med seboj zelo razlikujejo, zato se razlikujejo tudi oznake
- Nekaj truda, da so si vseeno čim bolj podobne:
 - angleške (sic!)
 - strukturne: <text>, <p>, <s>
 - pozicijske: word, (norm), lemma, tag, tag-en
- Vendar so atributi besedil zelo raznovrstni
- Nekateri korpusi imajo še nadaljnje pozicijske oznake (Universal Dependencies)
- Če tako nastavljeno v konfiguraciji, imajo oznake lahko tudi po več vrednosti.

Zaključki

Zaključki

- Kratka predstavitev konkordančnikov CLARIN.SI
- Večina korpusov, ki so dostopni na konkordančnikih je dostopnih tudi za prevzem v repozitoriju CLARIN.SI
- Nadaljnje delo:
 - Posodobiti instalacijo noSkE Crystal (lokalizacija)
 - Dodati še en konkordančnik: Crystal s prijavo
 - Nov sistem dodajanja novih korpus na konkordačnike
 - Tečaj(i) uporabe konkordančnikov
- Sedaj pa v živo!

Uporaba konkordančnika noSketch Engine @ CLARIN.SI

Tomaž Erjavec

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

Delavnica

"Uporaba digitalizirane zbirke pregovorov Pregovori 1.0 na CLARIN.SI"
2023-03-15
ZRC SAZU