Raziskovalna infrastruktura CLARIN.SI in njen pomen za jezikoslovne študije

Tomaž Erjavec

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

Predavanje na doktorskem študiju 3. stopnje
Primerjalni študij idej in kultur
ZRC SAZU
2021-12-06

Pregled predavanja

Kaj je CLARIN(.SI)

Raziskovalne infrastrukture

Kaj je RI?

Naprave, podatki in storitve, ki jih znanstvena skupnost uporablja pri raziskovanju na svojem področju.

- ESFRI: European Strategy Forum on Research Infrastructures, 2002
- Razvojni načrti: 2006 (35 RI), 2008, ..., 2018 (55), 2021 (66)
- 22 RI je organiziranih kot ERIC (European RI Consortium = evropska pravna oseba)
- Slovenija sodeluje v 20/22 RI (npr. CESSDA / ADP), 2 s področja humanistike:
- DARIAH ERIC / DARIAH-SI = Digital Research Infrastructure for the Arts and Humanities / Digitalna raziskovalna infrastruktura za umetnost in humanistiko: INZ + ZRC SAZU
- **CLARIN ERIC / CLARIN.SI** = Common Language Resources and Technology Infrastructure / Infrastruktura za jezikovne vire in tehnologije

CLARIN: 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 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 ERIC

- Sedež na Nizozemskem
- Trenutno 22 držav članic + 3 opazovalke
- Podporno osebje, odbori za upravljanje, delovne skupine
- Večina dela se odvija v okviru nacionalnih konzorcijev

Kaj CLARIN ERIC ponuja slovenskim raziskovalcem?

- S slovensko EduGain prijavo dostop do vseh virov in storitev centrov CLARIN držav članic
- Spletne storitve, npr. virtualni jezikovni observatorij
- Podpora ciljnim projektom, npr. razvoju učnih vsebin, izvedbi delavnic za uporabnike, snovanju evropskih projektnih prijav
- Infrastruktura znanja:

CLARIN.SI

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

- Začetek dela v 2014
- Institut "Jožef Stefan":
  - Odsek za tehnologije znanja (E8)
  - Laboratorij za umetno inteligenco (E3)
  - Center za mrežno infrastrukturo (CMI)
- Organiziran kot konzorcij 12 partnerjev:
  - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - 4 raziskovalni inštituti: ZRC SAZU, IJS, INZ, ZRS Koper
  - 2 društvi oz. zavoda: SDJT, Trojina
  - 2 podjetji: Amebis, Alpineon

Storitve CLARIN.SI

Trije stebri CLARIN.SI

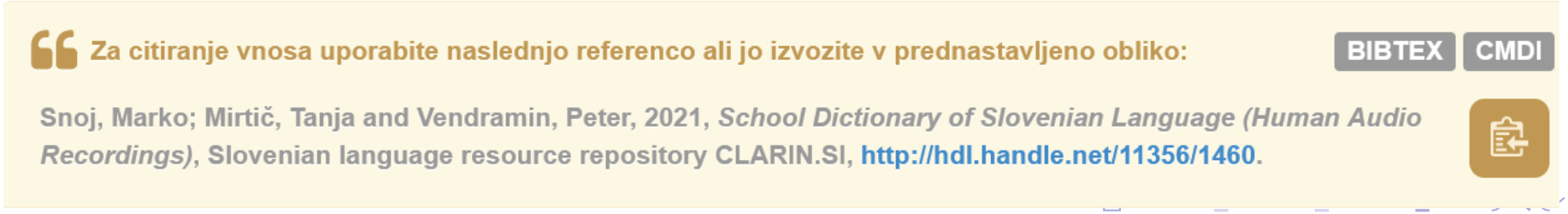
1. Repozitorij jezikovnih virov (in orodij)
2. **Dva konkordančnika** in druge spletne storitve
3. **Podpora (vsebinska in finančna)**

Repozitorij

- Zaenkrat najpomembnejša storitev CLARIN-a
- Stalna in varna hramba jezikovnih virov (certifikacija CTS)
- Pogoji uporabe (ToS, licence), etični kodeks (CoCo)
- Žetev metapodatkov (metadata harvesting)
- Trenutno vsebuje 254 virov, 173 (tudi) slovenskih
- Uporaba podatkov zahteva nekaj računalniškega znanja

Zakaj deponirati jezikovne vire

- Postopek objave vira:
  - avtor sam vnese metapodatke in naloži podatke (natančna dokumentacija zahtev + za začetnike nudimo pomoč)
  - urednik pregleda vnešeno in objavi
  - ali zahteva popravke
  - delno izpolnjen vnos je mogoče prenesti drugemu (soavtorju) v izpolnjevanje
- Vir zavarujemo pred izginotjem
- Naredimo korak k odprti znanosti
- Večja možnost, da vir najdejo in uporabijo drugi, in ga citirajo (stalni identifikatorji!):



Podpora

- CLARIN.SI organizacijsko ali finančno podpira dogodke v Sloveniji s področij (digitalnega) jezikoslovja in jezikovnih tehnologij, npr. konference "Jezikovne tehnologije in digitalna humanistika" (1998, . . . , 2020, . . .)
- Podpora razvoju jezikovnih virov in orodij za slovenščino:
  - podpora pripravi virov za vključitev v repozitorij. cca. 500 EUR
  - večji projekti: od 2018, letno 30.000 EUR za cca. 5 projektov

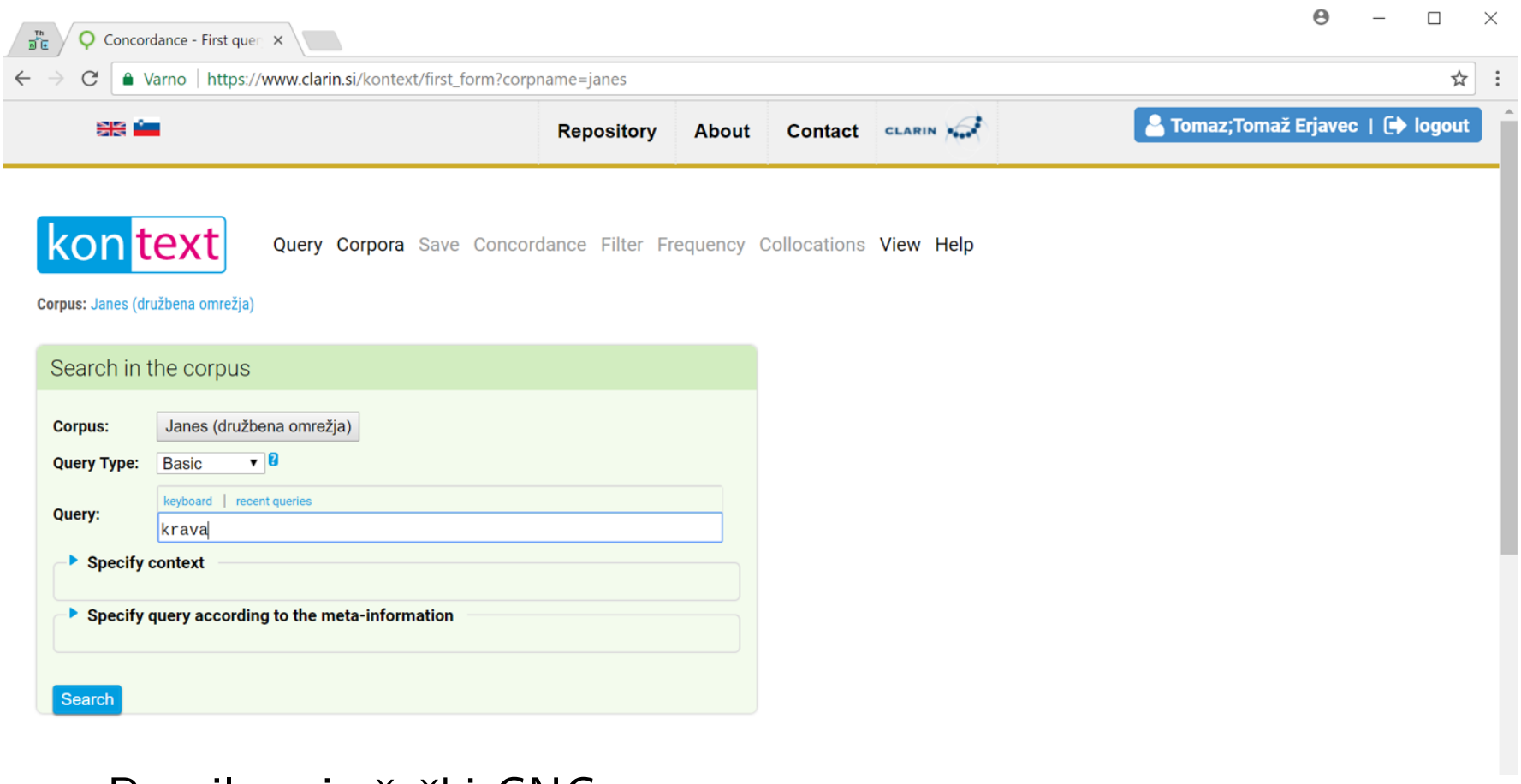
CLASSLA: Center znanja za južnoslovanske jezike

- CLASSLA CLARIN K-Centre for South Slavic Languages: CLARIN.SI + CLADA (2019)
- Strokovna podpora pri uporabi jezikovnih virov in tehnologij za južnoslovanske jezike
- Posredovanje informacij prek dokumentacije o razpoložljivih virih in tehnologijah raziskovalcem, študentom in drugim zainteresiranim posameznikom
- Tehnična podpora pri ustvarjanju, preoblikovanju in objavi jezikovnih virov in tehnologij
- Organizacija izobraževanj
- FAQ za slovenščino, hrvaščino, srbščino

Konkordančnika

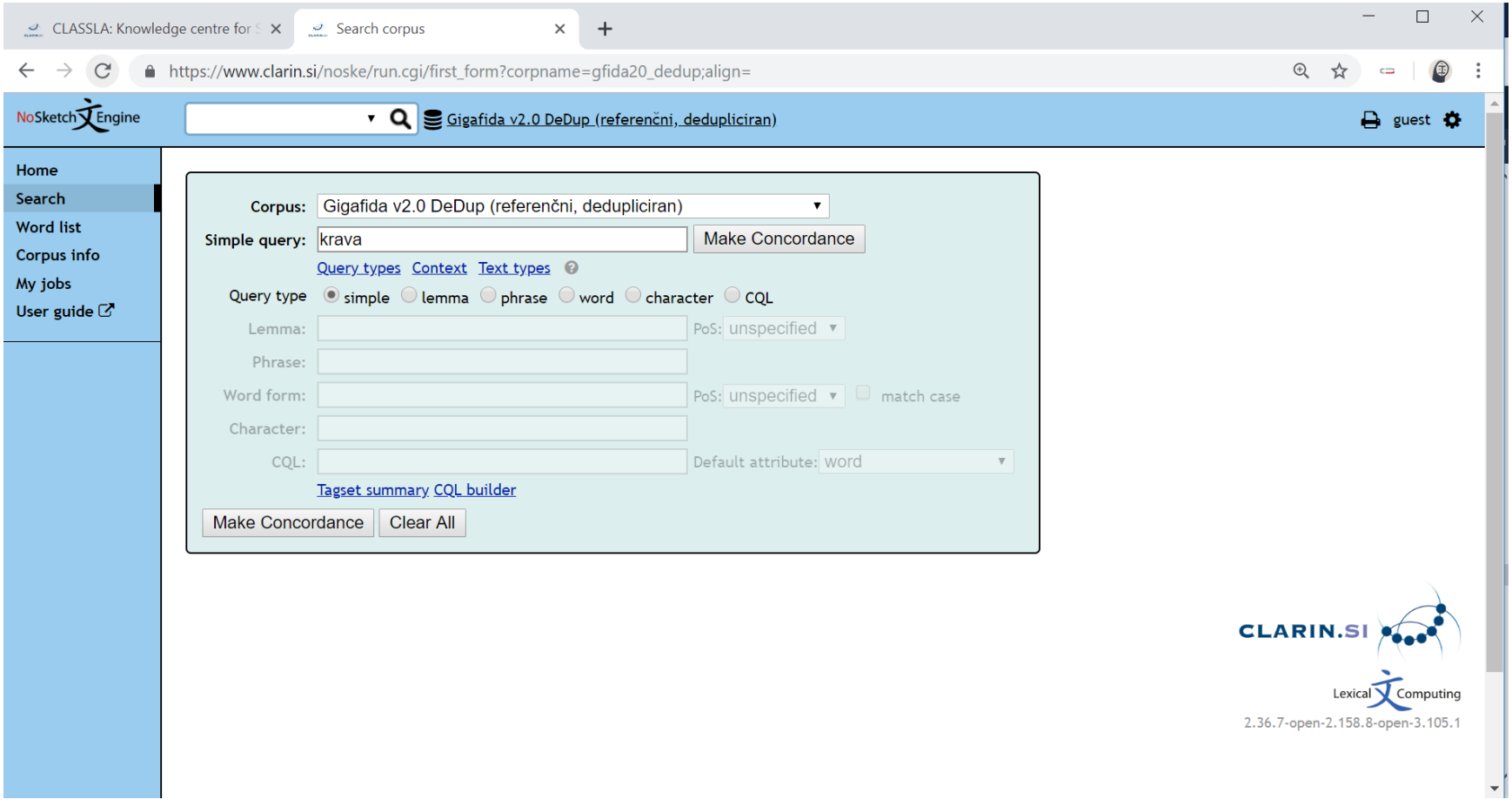
- KonText + noSketch Engine @CLARIN.SI
- Uporabljata isti zaledni program: Manatee
- Delata lahko z velikimi korpusi (več milijard besed)
- Korpusi so lahko bogato označeni:
  - strukture (besedilo, odstavek, termin, . . .)
  - metapodatki (datum, vrsta besedila, spol avtorja,. . .)
  - atributi pojavnic (oblikoskladenjska oznaka, lema, normalizirana oblika,. . .)
- Bogat poizvedovalni jezik: CQL
- Raznovrstni izpisi in analize
- RESTful, tj. poizvedbe prek URLjev (TXT, JSON ali XML)
- Ponujata cca. 100 korpusov v 33 jezikih z 20 milijard besed

KonText



- Razvil ga je češki CNC
- Prijava: shranjene poizvedbe, nastavitve zaslona, podkorpusi
- Nima nekaterih funkcionalnosti noSketch Engine

noSketch Engine



- Odprtokodna različica komercialnega Sketch Engine
- Prijava ni niti potrebna niti mogoča

Kaj konkordančnika omogočata

- Iskanja:
  - Iskanje besed, fraz ali delov besed
  - Iskanje po pojavnici, lemi, oblikoskladenjski oznaki...
  - Omejitev iskanja glede na lastnosti besedil (ali drugih strukturnih elementov)
- Izpisi:
  - Konkordance
  - Različne vrste frekvenčnih seznamov
  - Kolokacije
  - Število pojavitev po zvrsteh besedil
- Ostalo:
  - Nastavitve pogleda
  - Razvrščanje konkordanc
  - Shranjevanje podkorpusv
  - Izvoz rezultatov

MetaFida

Motivacija

- Korpusi na konkordančnikih CLARIN.SI so zelo raznovrstni, npr.
  - Referenčni: GigaFida
  - Znanstvena besedila: KAS
  - Starejša slovenščina: IMP
  - Uporabniško generirane vsebine: Janes
- Uporabniki bi radi iskali ali primerjali rezulate nekega iskanja po več korpusih, vendar:
  - je zamudno izvajati enaka poizvedovanja po več korpusih
  - hitro lahko pride do napak
  - zaradi raznovrstne strukture posameznih korpusov primerjave velikokrat sploh niso možne

MetaFida

- Izdelava zrduženega korpusa slovenskih korpusov na konkordančnikih (projekt RSDO)
- Izbira korpusov (=34 korpusov):
  - Čim večje število čim bolj raznovrstnih korpusov
  - Vključeni taki, ki niso podmnožica drugih korpusov
  - Pojavnice označene vsaj z lemo in oblikoskladenjsko oznako
- Pretvorba korpusov:
  - Identifikacija strukturnih in pozicijskih oznak
  - Pretvorba v nabor oznak MetaFida
  - Obdržimo samo tiste oznake, ki so skupne več korpusom
- 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 (novi korpusi, mogoče mehka deduplikacija)

Velikost in zgradba MetaFide

Pojavnic	4.463.244.126
Besed	3.646.106.520
Stavkov	228.797.305
Odstavkov	89.596.744
Besedil	15.338.751

- Strukturne oznake:
  text: corpus_id, corpus, info, id, year (-19.4 %), publisher (-3.8 %), title (-75 %), author (-99.4 %);
  p: id; **s**; **gap**; (**g**)
- Pozicijske oznake: **word**; **norm**; **lemma**; **tag_en**; **tag**
  + dinamični atributi lc; norm_lc; lemma_lc

Primer iz 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
```

Primer poizvedbe

NoSketch Engine | kostanj | metaFida v0.1 (združeni korpus)

Home
Išči
Seznam besed
O korpusu
My jobs
User guide

Shrani
Make subcorpus
Možnosti prikaza
  Usredinjeno
  Stavek
Razvrščanje
  po levi
  po desni
  iskani niz
  Podatki
  Premešaj
Vzorec
Filter
  Sub-hits
  1. zadetek v dokumentu
Frekvence
  Oznake niza
  Oblike niza

Iskalni niz kostanj 29,329 (6.57 na milijon)

Stran 1 od 978 Pojdi Naslednja | Zadnja

imp,WIKI00...	dobru sadene. O kreſu, kir je veliku	Kostajna	, je dobru stare Payni pomladiti,
imp,NUK_13...	, mandelne, leſhenke, orę̃he,	koſtan	, v'eno s'dobro poprej malu
imp,WIKI00...	. One jo nosijo is popovja divjih	kostanjov	, topolja ino drugega drevja. ɡ 72. K'
imp,WIKI00...	jablani, gruške, slive, tudi laški	kostanji	, ino proti konci totega meseca tudi
imp,WIKI00...	ſe. ɡ Sa lipo saſlushi predragi	koſtanj	pervo méſto, in she sato, ko to drevo
imp,WIKI00...	in gole proſtora s' divjim	koſtanjem	, kteri tudi na nar puſteji semlji
imp,NUKP14...	, snano je, de v kebrovim leti hraſt,	koſtanj	, oreh ino druge dreveſa bliso do
imp,NUKP14...	; per léſki, oréhu, hraſtu in	koſtanju	ſo tako imenovani brenklji ali
imp,NUKP14...	tovorizhi, kar je nam snano , ſo is	koſtanja	in nar vezhi is ſlibovne. ɡ
imp,NUKP14...	preš ali pa s posoljeno moko divjiga	kostanja	. ɡ Marnja. ɡ Na dveh sosednih njivah
imp,FPG_00...	kolerabami. ɡ Višnjev ohrovt s	kostanjem	. ɡ Peresa od štoržev odberi, čisto
imp,FPG_00...	, perdeni pečeniga in olušeniga	kostanja	, de se dobro skuha. ɡ Špargelnov
imp,FPG_00...	Tudi lahko popra in španske čebule,	kostanja	ali krompirja vanjo deneš in jo tako
imp,FPG_00...	drobnih, mandelnov in pečeniga	kostanja	vzame, tedaj se mora pa šest lotov na
imp,FPG_00...	na taljarčku in daj na mizo. ɡ Cukreni	kostanj	. ɡ Lepiga debeliga kostanja speci in
imp,FPG_00...	. ɡ Cukreni kostanj. ɡ Lepiga debeliga	kostanja	speci in olupi. Potem pa zavri cukra
imp,FPG_00...	. Vari, de cuker rujav ne postane.	Kostanj	pa na igle natakni, nekolikrat v
imp,FPG_00...	goveje mesa. ɡ Višnjev ohrovt s	kostanjem	in mesenimi klobasicami brez čev. ɡ
imp,NUKP14...	njih, bore, smreke, hraste, orehe,	kostanje	, jesene in breste izrediti, pri
imp,NUKP14...	driska napadla, stolci divjiga	kostanja	ali še bolje želoda v moko, in daj mu 2
imp,NUKP14...	to se po pravici mandeljni, orehi in	kostanj	prištevajo k redivni hrani. Mnoge

Primer sortiranja

NoSketch Engine | kostanj | metaFida v0.1 (združeni korpus)

Home
Išči
Seznam besed
O korpusu
My jobs
User guide

Shrani
Make subcorpus
Možnosti prikaza
  Usredinjeno
  Stavek
Razvrščanje
  po levi
  po desni
  iskani niz
  Podatki
  Premešaj
Vzorec
Filter
  Sub-hits
  1. zadetek v dokumentu
Frekvence
  Oznake niza
  Oblike niza

Iskalni niz kostanj 29,329 > Multilevel Sort 29,329 (6.57 na milijon)

Prva | Prejšnja Stran 362 od 978 Pojdi Naslednja | Zadnja Konkordance so razvrščene. Pojdi na: gos11

gfida20_de...	masla ali masti. Po nadevu potresi	kostanj	, ki mu primešaj za veliko prgišče
gfida20_de...	pošteno potruditi pri nabiranju	kostanja	, ki ga letos ni bilo ravno v izobilju
gfida20_de...	pomagali pri organizaciji in peki	kostanja	ter poskrbeli za prijetno druženje
gfida20_de...	v lepi septembrski soboti podal po	kostanj	na Blegoš, Milan Vošank pa po
gfida20_de...	Dejmo Stisnt Teater, pečenega	kostanja	, kuhanega vina in toplega čaja. Pri
gfida20_de...	pa s povabilom na prvi jesenski	kostanj	ter topel napitek. Programski del
gfida20_de...	gospoda Christiana Zaichena iz	Kostanj	na avstrijskem Koroškem. On vsako
gfida20_de...	Kostanje I Prijazna ta vasica je	Kostanje	, očaran vsak nad njeno je lepoto;
gfida20_de...	skupaj s koroškimi Slovenci iz	Kostanj	(Kőstenberg - Avstrija) na Triglav
gfida20_de...	, je letos obrodilo skupen pohod od	Kostanj	do Triglava. Prvo srečanje je bilo
gfida20_de...	strnil Christian Zeichen iz	Kostanj	na Avstrijskem Koroškem, ki je
gfida20_de...	borovnic, jeseni pa je precej tudi	kostanja	. Sredi gozdička nas je presenetil
gfida20_de...	gorami in petimi gozdovi je živel	kostanj	Kostanjček. Njegova starša sta
gfida20_de...	načrtujemo večdnevni pohod iz	Kostanj	(Avstrija) v smeri: Vrbsko jezero-
gfida20_de...	kaskada, ljubka jezerca, ogromni	kostanji	. Rozarij, velik vrt z
gfida20_de...	ponudili še krompirjeve svaljke s	kostanjem	v drobljencu vijoličastega
gos11,gos1...	greva kr na kostanjevo strjenko ne? ɠ	kostanji	ɠ ja evo ɠ upam da jih ne bo treba lupet ɠ ne
gos11,gos1...	kr pr pretresla ɠ če se ti razkuhajo	kostanji	je v bistvu dobr kr ti pol razpadejo
gos11,gos1...	sem bolj malo hodo rajš sem hodo po	kostanj	ker se je dobro prodajal ɠ a no vidiš
gos11,gos2...	pol vstanejo grejo malo nabirat	kostanje	je bla lih una štajon od kostanjev ne ɠ
gos11,gos2...	kostanje je bla lih una štajon od	kostanjev	ne ɠ ratajo lačni ɠ razumeš grejo malo

Zaključki

Zaključki

Kaj lahko CLARIN(.SI) ponudi jezikoslovcem:

- CLARIN ERIC omogoča dostop do:
  - storitev centrov CLARIN: VLO (repozitoriji), orodja
  - znanja (K-centres / centri znanja)
  - sredstev: sodelovanje na konferencah CLARIN, organizacija konferenc, izdelava učnih gradiv itd.
- CLARIN.SI omogoča dostop do:
  - repozitorija (prevzem, deponiranje)
  - konkordančnikov (in drugih spletnih storitev)
  - znanja: CLARIN.SI pomoč, CLASSA center znanja
  - sredstev: vključitev virov v repozitorij, projekti CLARIN.SI
- Novi korpus MetaFida v0.1: če najdete napake ali imate predloge za izboljšave, prosim sporočite!
- Vse, kar je bilo danes predstavljeno, je podrobneje opisano na spletiščih:
  - `https://www.clarin.eu/`
  - `https://www.clarin.si/`

Raziskovalna infrastruktura CLARIN.SI in njen pomen za jezikoslovne študije

Tomaž Erjavec

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

Predavanje na doktorskem študiju 3. stopnje
Primerjalni študij idej in kultur
ZRC SAZU
2021-12-06