Korpusi slovenskega jezika: dostopnost, oznake in zapis

Tomaž Erjavec, IJS

ZRC SAZU, 7. 12. 2017

Pregled predavanja

Uvod

Pomembnost korpusov

- Slovaropisje
- (Korpusno) jezikoslovje
- Poučevanje jezika
- Prevodoslovje
- Jezikovne tehnologije

Kvalitete korpusov

Priporočila EAGLES, J. Sinclair (1996):

- *avtentičnost*: korpus ustreza kriterijem, glede na katere je bil narejen
- *velikost*: čim večji, tem boljši
- *kakovost*: zapis in oznake korpusa so pravilne
- *enostavnost*: računalniški zapis korpusa je razumljiv
- *dokumentiranost*: korpus je opremljen z bibliografskimi in drugimi metapodatki

Nekateri korpusi slovenskega jezika

- **GigaFida** (1.000 mil. besed: reprezentativni (kmalu v1.1 + dedup, v2.0)
- **Gos**, **GosVL 2.0** (25 predavanj, 80,000 besed, 10 ur): govorni
- **IMP** (658 besedil, 25 mil. besed): starejša slovenščina
- **Janes** (200 mil. besed): slovenščina družbenih medijev
- **slWaC** (750 mil. besed): slovenski splet
- **KAS** (780 mil. besed): zaključna dela s slv. univerz
- **slovParl v2.0** (11 mil. besed): parl. razprave 1990-1992
- **JaSlo**, **LeMonde**, **TRANS5**, **DGT15**: vzporedni
- **Šolar**, **Lektor**, **KoRP** itd.

(tudi več ročno označenih korpusov)

Dostopnost

Dostopnost in odprtost korpusov

Miran Hladnik, Nagovor ob podelitvi Trubarjevega priznanja 2017:

Ljubi rojaki, veste, kaj letošnji "varuh kulturne (pisne) dediščine"napravim, ko naletim nanjo? Najprej jo fotografiram, besedila po možnosti prepišem [...], opremim z licenco proste dostopnosti in uporabnosti in postavim na javno spletno mesto, kjer si lahko vsakdo, kadarkoli, zastonj in za kakršen koli namen postreže z njo. Temu ne bi ravno rekel varovanje, [...] nasploh mi je brambovski ali zaščitniški odnos do kulturnih dobrin tuj in vzdrževanje njihove nedotakljivosti nesprejemljivo. Če hočemo biti kulturni, moramo kulturo uporabljati, jo živeti vsak dan, ne le za nedelje in praznike.

Ovire

- Pravice in omejitve nad izvornimi besedili, *ibid*:
  Veste, kaj me ovira pri početju, ki sem ga opisal? Slovenska zakonodaja, konkretno Zakon o varstvu kulturne dediščine, Zakon o avtorskih pravicah, Zakon o varovanju zasebnosti, Zakon o dostopu do informacij javnega značaja, Zakon o varstvu dokumentarnega in arhivskega gradiva.
- Avtorstvo samega korpusa

Rešitve

- Načelna zavzetost za jezik kot javno dobro
- Kjer je mogoče: dobiti dovoljenja
- Kjer se tako presodi: anonimizirati, premešati, vzorčiti
- Nekonfliktno in hitro upoštevanje pritožb, tj. izbris spornih vsebin
- Tehnični vidiki: principi FAIR (Findable, Accessible, Interoperable, Reusable)

Načini dostopa

- Konkordančniki
- Druga spletna analitična orodja
- Digitalne knjižnice: natančno branje (*close reading*)
- Povezovanje:
  - korpus → digitalna knjižnica
  - slovar → korpus
- Prevzem:
  - poljubne in poglobljene analize
  - učenje, testiranje in uporaba jezikovnotehnoloških orodij
  - CC: ponovna uporaba, izboljšave, nadgradnja

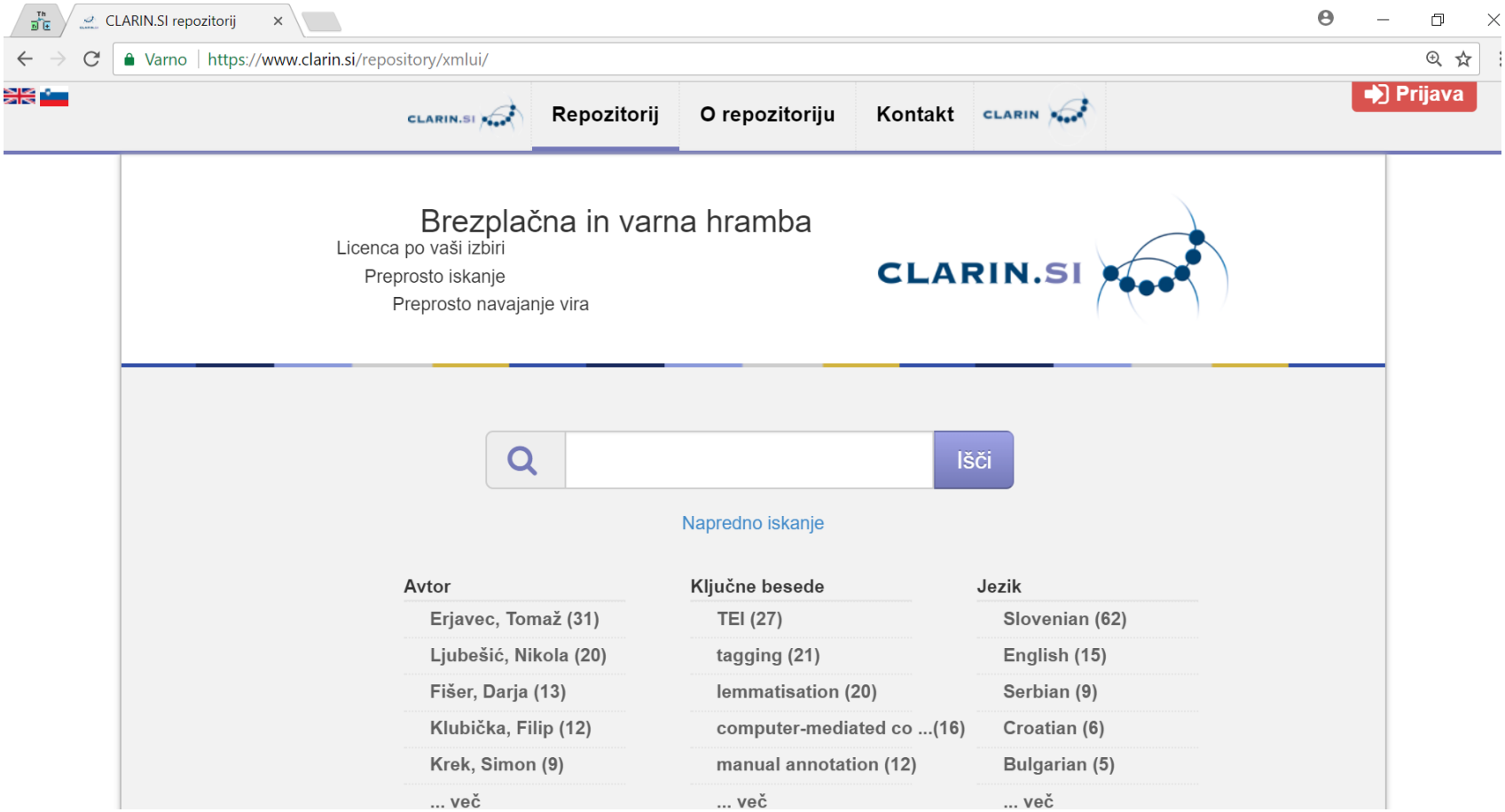
CLARIN.SI

- Raziskovalna infrastruktura za jezikovne vire in tehnologije
- Članica Evropske infrastrukture CLARIN
- Konzorcij 12 partnerjev
- Storitve:
  - repozitorij jezikovnih virov: dSpace
  - konkordačniki: KonText (novo!), noSketch Engine
  - ročno označevanje besedil: WebAnno
  - avtomatsko označevanje besedil: WebLicht (v delu)

Repozitorij CLARIN.SI

- Trajno in varno arhiviranje korpusov in drugih jezikovnih virov
- Bogati metapodatki + priporočeno citiranje (handle)
- Izbira licence
- Prijava prek AAI (EduGain) + GÉANT CoCo (spremljanje uporabe)
- Žetev metapodatkov
  (CLARIN VLO, OpenArchives Re3Data, OpenAIRE, ...)
- Lokaliziran uporabniški vmesnik (ne pa metapodatki!)

Repozitorij



KonText in noSketch Engine @ CLARIN.SI

- Novo: konkordačnika sedaj v sklopu CLARIN.SI
- Vsak s svojimi prednostmi (in slabostmi)
- Ponujata isti nabor korpusov (trenutno okoli 40)
- Povezava z repozitorijem
- Lokaliziran uporabniški vmesnik (v delu)
- Omogočata zunanje povezave (npr. Slovar tviterščine)
- Enostaven prenos na druge instalacije (no)Sketch Engine ali KonText

Digitalne knjižnice

- Natančno branje
- Poskusi: eZISS, eZMono, TEITOK
- Prihodnost: DARIAH-SI

IMP
Digitalna knjižnica

EN | SL

Glavni meni

- Domača stran
- Iskanje
- Login

Powered by TEITOK
© Maarten Janssen, 2014

kpl-dipl

Kapelski pasijon

Nastavitve prikaza

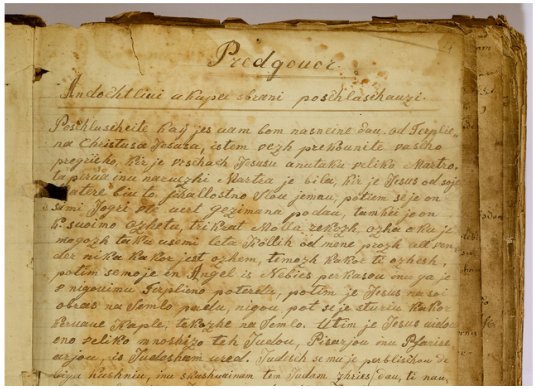
Besedilo: transkripcija | normalizirano - Prikaži: barve | Formatting | <lb> | faksimile - Oznake: lema

indeks 000r < Folio 001 > 002

1 Prvi del Predgouor

Andochtliui ukupei sbrani poschluschauzi.

Poschluscheite kaÿ jes uam bom nasneine dau. od Terplie=
na Christusa Jesusa, istem vezh prekounite vascho
pregrieho, kir je vrschach Jesusu anutaku veliko Martro,
ta perua inu naruezhi Martra je bila, kir je Jesus od soje
Matere biu to schallostno Slou jemau, potiem se je on
is soimi Jogri vte uert Gezimana podau, tamkei je on
ksuoimo Ozhetu, trikrat Mollu rekozh, ozha aku je
mogozh taku usemi leta Köllih od mene prozh all ven-



Oznake

Vrste oznak

- Metapodatki:
  - avtor, naslov datum, založba, ... besedila
  - spol, starost, ... avtorja
  - standardnost, sentiment, ... besedila
- Jezikoslovne oznake:
  - posamezne besede: lema, oblikoskladenjska oznaka
  - nizi besed: imena, večbesedne enote
  - medsebojne povezave: odvisnostna skladnja
  - zunanje povezave: slovarji, enciklopedije, ontologije

Označevanje

- Razen za zelo majhne korpuse ročno označevanje ni možno
- Avtomatsko označevanje:
  - učna množica + strojno učenje = model
  - program + model = označevalnik
- Nobena avtomatska metoda ni popolna!
  iščemo npr. samostalnike
  - *natančnost*: program označi kot samostalnik tudi nekatere besede, ki to niso
  - *priklic*: program ne označi kot samostalnik nekatere besede, ki so samostalnik
- Veriženje orodij: množenje napak
- Uporaba različnih orodij: problemi s primerjavami

Označevanje slovenskega jezika

`"Da se naujo zdaj še na Planico spravl!?"`

Obstoječi moduli:

- **Tokenizacija in stavčna segmentacija**
  `[Da|se|naujo|zdaj|še|na|Planico|spravl|!?]`
- **Normalizacija** (posodabljanje in standardizacija)
  `[da|se|ne|bojo|zdaj|še|na|planico|spravili|!?]`
- **Oblikoskladenjsko označevanje**
  `[Vd|Zp------k|L|Gp-ptm-n|Rsn|L|Dt|Slzet|Ggdd-mm|U]`
- **Lematizacija**
  `[da|se|ne|biti|zdaj|še|na|Planica|spraviti|!?]`
- **Skladenjsko označevanje**
- **Označevanje in kategorizacija imen**
  `<name type="geo">Planica</name>`

Zapis

Prednosti standardiziranih zapisov in oznak

- **Preverljivost**: skladnost s standardom je mogoče avtomatsko preveriti
- **Trajnost**: standard je javen, vzdrževan in dokumentiran
- **Interoperabilnost**: podatki niso vezani na programsko opremo, podatki so združljivi
- **Ponovna uporaba**: podatki se lahko uporabijo tudi za namene, za katere niso bili izvorno predvideni

Standardi za (CLARIN.SI) korpuse

Zapis:

- Unikod, XML: osnova
- TEI: Text Encoding Inititative: strukturne oznake in atributi
- ISO: datumi, kode jezikov

Jezikoslovje:

- ReLDI/SSJ: 2 x tokenizacija
- MULTEXT-East: oblikoskladenjske oznake, lematizacija
- JOS/SSJ: odvisnostna skladnja
- Universal Dependencies (!)
- Janes: imena

Primer

```
<choice>
    <orig>
        <w>vosi</w>
    </orig>
    <reg>
        <w lemma="voza" ana="#Ncf">vozi</w>
        <desc>
            <gloss>ječa</gloss>
            <bibl>[sskj]</bibl>
        </desc>
    </reg>
</choice>
<c> </c>
<name type="deriv-per">
    <choice>
        <orig>
            <w>Criſtuſeva</w>
        </orig>
        <reg>
            <w lemma="kristusov" ana="#Asp">kristusova</w>
        </reg>
    </choice>
</name>
```

Zaključki

Povzetek

Slovenščina ima že vrsto

- dostopnih, označenih in standardno kodiranih FAIR korpusov (CLARIN.SI repozitorij);
- spletnih storitev za analizo (KonText, noSketchEngine) in označevanje (WebLicht)

Nadaljnje delo

- Gradnja več in večjih korpusov
- Nadgradnja konkordančikov (analitika)
- Boljše osnovno označevanje in nove ravni označevanja
  ← večji in bolj raznovrstni ročno označeni učni korpusi in zaledni viri
- Povezovanje korpusov/konkordačnikov, digitalnih knjižnic in slovarjev
- Vzdrževanje specifikacij za jezikoslovno označevanje
  ← MULTEXT-East V5 / UD

metaFida

oz. Slovenski nacionalni korpus

- Združiti vse obstoječe korpuse slovenskega jezika (deduplikacija)
- Zapolniti diahrone praznine: 1500–1600, 1600–1800, 1918–1991
- dLib.si
- Boljše označevanje
- Bogata dokumentacija