Jezikovni viri, odprta znanost in prihodnost slovenskega jezika v digitalni družbi

Tomaž Erjavec

Odsek za tehnologije znanja

Kolokviji na IJS
2019-10-23

Uvod

Malce zgodovine

- IJS E4, Peter Tancig: začetek dela na računalniški obdelavi slovenščine
- Vendar: informacijski sistem za Slovenijales skladišče, program za registracijo delovnega časa
- Generativnih modeli jezikove kompetence (Chomsky)
  - Čeprav: Žagar, Igor Ž., Tancig, Peter (1989). Računalniška analiza "napadov na JLA". Časopis za kritiko znanosti (17) 119/120. - prvi korpus narejen na E4
- Frustracija: modeli narejeni za angleščino, ne najbolj primerni za slovenščino; nobene uporabne vrednosti
- Doktorska disertacija, 1997: "Unifikacija, nasledstvene hierarhije in paradigme v formalizaciji morfologije jezikov"

Prvi projekti EU

- 1995: EU Copernicus, prvič tudi za bivše soc. države
- Iz generative (kompetenca) v empirijo (performanca): bistveno bolj uporabno in smiselno za slovenščino
- Copernicus project MULTEXT-East "Multilingual Text Tools and Corpora for Central and Eastern European Languages"
- Copernicus concerted action TELRI "Trans-European Language Resources Infrastructure"
- Nikoli ne odnehaj: MULTEXT-East V2 (2002), V3 (2004), V4 (2010), V5 (2016), V6 (TBA)

George Orwell
Nineteen Eighty-Four

It was an enormous pyramidal structure of glittering white concrete, soaring up, terrace after terrace, 300 metres into the air. From where **Winston** stood it was just possible to read, picked out on its white face in elegant lettering, the three slogans of the **Party**:

War is peace	**Freedom is slavery**	**Ignorance is strength**
Războiul este pace	Libertatea este sclavie	Ignoranța este putere
Vojna je mir	Svoboda je suženjstvo	Nevednost je moč
Válka je mír	Svoboda je otroctví	Nevědomost je síla
Войната е мир	Свободата е робство	Невежеството е сила
Sõda on rahu	Vabadus on orjus	Teadmatus on jõud
Rat je mir	Sloboda je ropstvo	Neznanje je moć
A háború: béke	A szabadság: szolgaság	A tudatlanság: erő
Karas — tai taika	Laisvė — tai vergija	Nežinomas — tai jėga
Rat je mir	Sloboda je ropstvo	Neznanje je moć
Война — это мир	Свобода — это рабство	Незнание — сила

Prvi slovenski projekti izdelave jezikovnih virov

- FIDA, prvi referenčni korpus slovenskega jezika (1996)
- ARRS JOS "Jezikoslovno označevanje slovenskega jezika: metode in viri" (2007-2009)
- ESS + MIZŠ: SSJ "Sporazumevanje v slovenskem jeziku" (2007-2013)
- Nikoli ne odnehaj 2:
  - FidaPLUS (2006), GigaFida, (2009), GigaFida 2.0 (2019)
  - jos100k (2010), ssj500k v1 (2014), ssj500k v2 (2019)

Jezikovni viri

Kaj so jezikovni viri

- Pisni in govorni korpusi jezika: zbirke besedil
  - izbrane po vnaprej določenih kriterijih
  - opremljene z metapodatki
  - jezikovno označene
  - enovito kodirane
- Digitalni slovarji
- Računalniški leksikoni
- Modeli jezika

Uporabnost jezikovnih virov

- Humanistika:
  - jezikoslovje: slovaropisje, sociolingvistika, poučevanje jezika...
  - digitalni slovarji
  - veliki, avtomatsko označeni in standardno zapisani korpusi jezika
- Računalništvo:
  - računalniško procesiranje jezika: učne in testne množice
  - ročno označena besedila (nadzorovano strojno učenje)
  - podporni viri: leksikoni
- Različnost kultur in usmeritev humanistike in naravoslovja
  - slovenščina / angleščina
  - lokalno / globalno
  - mehko / trdo
  - zaprto / odprto

Delotok izgradnje korpusa

1. Zbiranje besedil
2. Oblikovanje v korpus (izbira, metapodatki, zapis)
3. Priprava avtomatskega označevanja:
   1. izdelava smernic za označevanje
   2. ročno označevanje vzorca
   3. učenje modela na ročno označenem vzorcu
   4. evalvacija
4. Označevanje celotnega korpusa
5. Priprava za objavo

Primer JANES

- ARRS JANES "Jezikoslovna analiza nestandardne slovenščine", 2014–2018 (Darja Fišer, FF)
- Prvi veliki korpus slovenskih uporabniško ustvarjenih besedil: tviti, blogi, komentarji
- Izdelava:
  1. zbiranje besedil: zajem prek Twitter API + zajem s spleta
  2. oblikovanje v korpus: metapodatki o viru, času, uporabniku, všečkanjih itd.
  3. avtomatsko označevanje:
     - standardizacija besed: "lohk" → "lahko"
     - oblikoskladenjsko označevanje in lematizacija: "Tej *hiši* manjka fasada" → "hiša" / "Ncmsd=Somed"
     - spol avtorja (ženski, moški, neznan)
     - standardnost besedil (L1-L3; T1-T3)
     - sentiment besedil (negativno, nevtralno, pozitivno)

Pomembnejši korpusi slovenščine

- Gigafida 1.0 (1.400M), Gigafida 2.0 (1.800M): referenčni
- slWaC (+hrWaC, srWaC, bsWaC): spletna besedila (900M)
- JANES: uporabniško ustvarjena besedila (250M)
- IMP: starejša besedila 1584-1919 (18M)
- KAS: zaključna dela s slovenskih univerz 2000-2018 (2.000M!)
- siParl: parlamentarne seje 1990-2018 (230M)
- GOS: govorjena slovenščina (1M)
- GosVL: predavanja z VideoLectures (0.18M)

Pomembnejši ročno označeni korpusi slovenščine

- ssj500k (500k): oblikoskladnja, lematizacija, skladnja, imena, udeleženjske vloge, glagoske večbesedne zveze
- goo300k (300k): posodabljanje, oblikoskladnja, lematizacija
- janesNorm (180k): standardizacija besed
- janesTag (75k): oblikoskladnja, lematizacija, imenske entitete

CLARIN.SI

Raziskovalne infrastrukture

Kaj je RI?

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

Raziskovane infrastrukture ESFRI

- *European Strategy Forum on Research Infrastructures* (ESFRI) je predlagal 15 (2016: 21) RI, nekatere že delujejo kot pravne osebe *European RI Consortium* (ERIC)
- Slovenija sodeluje v 14 ESFRI RI (npr. CERN)
- CLARIN ERIC / CLARIN.SI: Infrastruktura za skupne jezikovne vire in tehnologije (Common Language Resources and Technology Infrastructure)

CLARIN ERIC

- Sedež na Nizozemskem
- 20 držav članic + 4 opazovalke
- Podporno osebje, odbori za vodenje, delovne skupine
- Večina dela se odvija v okviru nacionalnih konzorcijev

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
  - 3 raziskovalni inštituti: ZRC SAZU, IJS, INZ
  - 3 društva oz. zavodi: SDJT, Trojina, DDR
  - 2 podjetji: Amebis, Alpineon
- Plodno sodelovanje s DARIAH-SI/INZ in CESSDA-SI/FDV

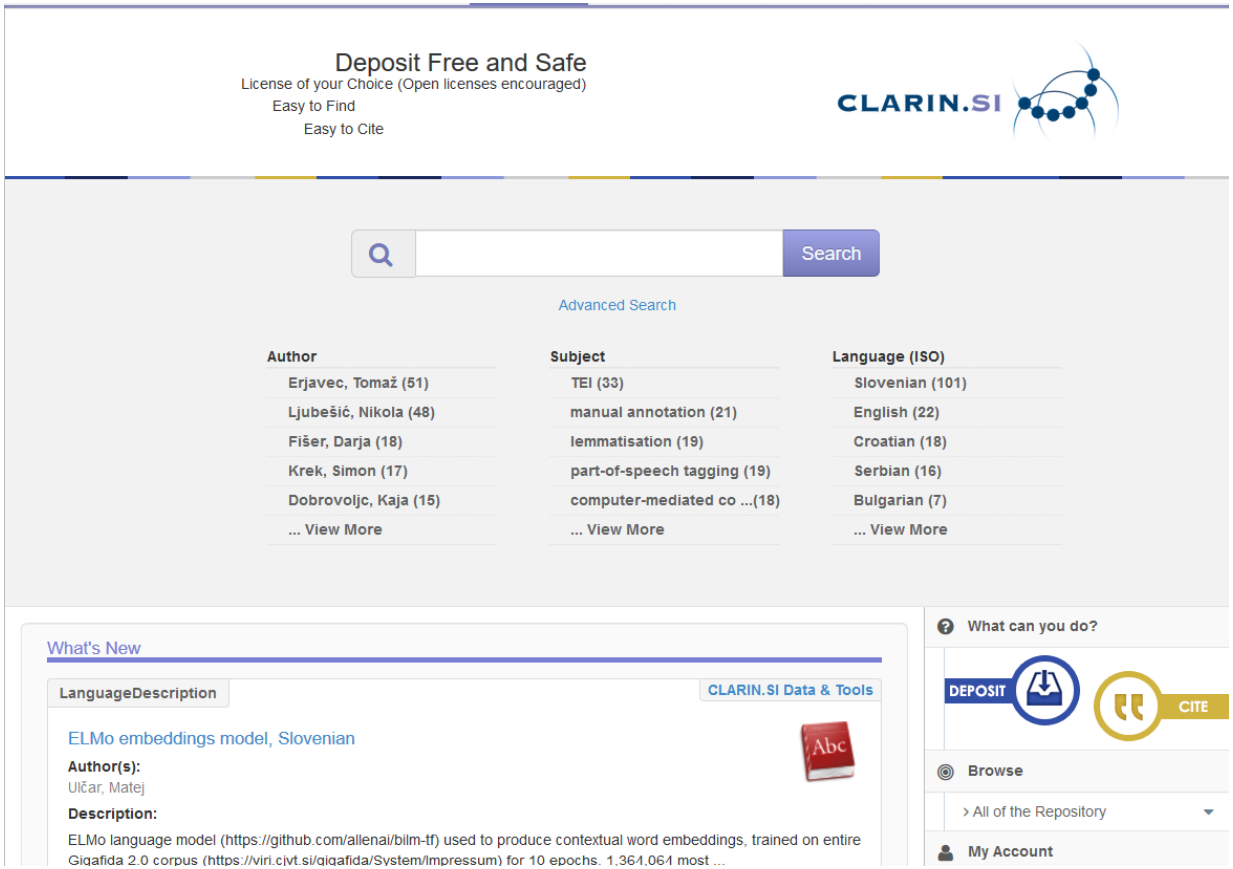
CLARIN.SI storitve

Trije stebri:

- Certificiran repozitorij jezikovnih virov in orodij
  - dolgotrajno hranjenje
  - avtentikacija in avtorizacija
  - stalni identifikatorji
  - eksplicitni pogoji uporabe in licence
  - principi FAIR
- Spletne storitve
  - dva konkordančnika (spletne analize korpusov)
  - orodja za označevanje besedil
- Podpora:
  - financiranje priprave virov za vključitev v repozitorij
  - večji projekti: 30.000 EUR letno: 7 projektov v 2018, 6 v 2019
  - dogodki: JOTA, JT-DH 2018, EURALEX 2018, TSD 2019, ...
  - CLASSLA K-centre: CLARIN.SI center znanja za računalniško obdelavo južnoslovanskih jezikov

Repozitorij

- Najpomembnejša storitev CLARIN.SI
- Danes 137+ jezikovnih virov, od tega 101 slovenskih: korpusi, slovarji, besedišča, modeli, programi
- Velika večina pod eno od licenc Creative Commons



Odprta znanost

Ponovna uporaba

Klasični pristop

- za vsako raziskavo posebej izdelati jezikovne vire
- viri dostopni samo razvijalcem

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

Odprti 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

Načela "FAIR"

- Findable, Accessible, Interchangeable, Reusable
- projekti EU za omogočanje odprtega dostopa: EOSC
- FACT: fair, accurate, confidential, transparent

Problemi pri zagotavljanju odprtega dostopa

- Vir obstaja, a ga je težko najti
- Vir nima opisa (metapodatkov), ali pa je le-ta pomanjkljiv
- Tehnične ovire: posebni in nedokumentirani formati zapisa
- Pravne:
  - avtorske pravice nad besedili
  - varovanje zasebnosti (tudi pravica do pozabe): GDPR
  - pogoji uporabe spletnih portalov (npr. Twitter)
- Več dela za izdelovalce virov!

Zgodnji trud za odprti dostop

- Strežnik nl.ijs.si deluje od 1994, trajni URLji
- MULTEXT-East, TELRI
- GNUsl (z Alešem Koširjem in Primožem Perterlinom): lokalizacija odprte kode (konec '90)
- Ustanovni predsednik SDJT (1998): eksplicitno prizadevanje za odprti dostop; vsi konferenčni zborniki na spletu
- Član Sveta TEI "Text Encoding Initiative Consortium" (2001 – 2002)

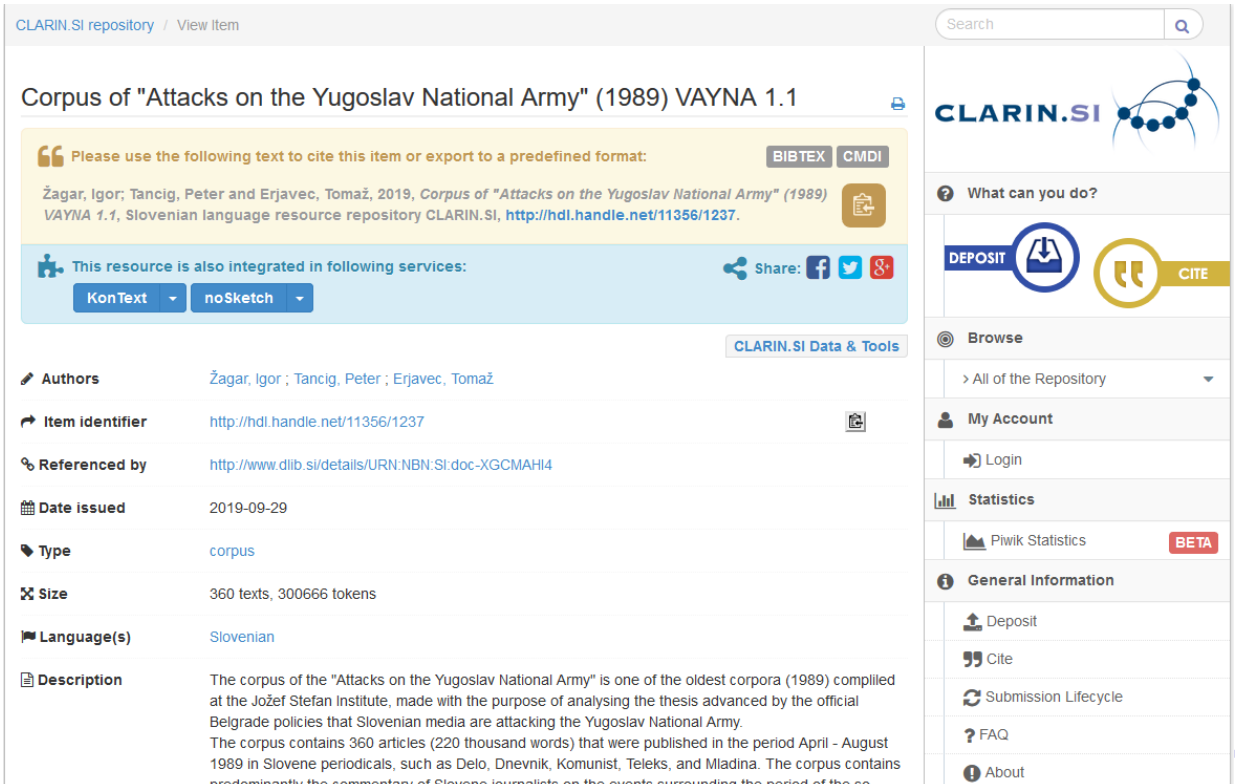
... in kasnejša prizadevanja

- Član skupin za pripravo:
  - Akcijski načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov financiranih z javnimi sredstvi v RS (2013)
  - Smernice za zajem, dolgotrajno ohranjanje in dostop do kulturne dediščine v digitalni obliki v RS (2013)
  - Nacionalni program za jezikovno politiko RS 2012 – 2016 (2012)
  - Akcijski načrt za uresničevanje Nacionalnega programa (2014)
- Nacionalni koordinator CLARIN.SI (2014-)
- Svetovalec na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (2019-)
- Sodelovanje v RDA Slovenija, vodja FDV (2019)
  - Delovna skupina za Koordinacijo infrastrukturnih podatkovnih storitev (ustanovni sestanek 14. 11. 2019, Maribor)

Zaključki

Povzetek

- Predstavil dologoletna prizadevanja za zagotovitiv odprtih jezikovnih virov za slovenski jezik
- Vsi predstavljeni viri narejeni v sodelovaju z drugimi raziskovalci
- Publiciranje FAIR in dolgotrajno hranjenje jezikovnih virov zelo olajšano z vzpostavitvijo CLARIN.SI



Prihodnost slovenskega jezika v digitalni družbi?

- Slovenščina je eden od uradnih jezikov EU: v bistveno boljšem položaju kot marsikateri jezik
- Aktivna raziskovalna skupnost: IJS, UL/CJVT, UM, ...
- Nacionalni in EU projekti za razvoj jezikovnih virov in tehnologij (tudi) za slovenščino
  npr. EMBEDDIA "Cross-Lingual Embeddings for Less-Represented Languages in European News Media" (Senja Pollak, E8)
- Odprt in brezplačen dostop do virov, orodij in storitev skozi CLARIN.SI

Prihodnje delo

- Poslanstvo
  - Odprt in brezplačen dostop do virov, orodij in storitev za slovenske raziskovalce in — kjer le mogoče — podjetja
  - n.b.: slovenski raziskovalci imajo dostop tudi do storitev CLARIN ERIC in drugih nacionalnih konzorcijev CLARIN
- Izzivi
  - avtorske pravice, varovanje zasebnosti
  - vrednotenje razvitih virov in tehnologij
  - navajanje uporabljenih virov in tehnologij v publikacijah
  - metodološka in tehnična znanja v raziskovalni skupnosti
- Priložnosti
  - digitalna humanistika in družboslovje
  - vse ostale vede, kjer je jezik pomemben
  - jezikovne tehnologije

Jezikovni viri, odprta znanost in prihodnost slovenskega jezika v digitalni družbi

Tomaž Erjavec

Odsek za tehnologije znanja

Kolokviji na IJS
2019-10-23