Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres

Simon Krek, Polona Gantar, Špela Arhar Holdt, Vojko Gorjanc

Projekt

- Ministrstvo za kulturo / Univerza v Ljubljani (CJVT)
- Pogodba: 4. 11. 2015
- Trajanje: 2015-2018
- Finance
  - 8.500 EUR na leto
  - skupaj 34.000 EUR
- Okvir
  - uresničevanje Resolucije o nacionalnem programu za jezikovno politiko 2014-2018, **nad/gradnja in vzdrževanje korpusov slovenščine**
  - spodbujanje razvoja jezikovnih tehnologij za slovenski jezik, ki vključuje vzpostavitev potrebne infrastrukture ter izdelavo čim bolj prosto dostopnih virov in orodij ter podatkov o slovenščini

Cilji

- usmerjeno **zbiranje** novih gradiv
  - (a) besedila, za katera je bilo glede na tip, zvrst ali druge kriterije po analizi korpusov Gigafida in Kres ugotovljeno, da so **podreprezentirana**
  - (b) besedila izbranih spletnih besedilodajalcev z večjo produkcijo, ki zagotavljajo večjo **aktualnost** korpusnega gradiva
- **strojna obdelava** novih (in obstoječih) gradiv
- **javna dostopnost** nadgrajenih korpusov, distribucija in diseminacija

- ciljni obseg: 1,5 milijarde besed

Usmerjeno zbiranje novih gradiv - 1

- **šolska** gradiva
  - (prosto dostopni) učbeniki, delovni zvezki ter sorodna učencem ter dijakom namenjena besedila vseh šolskih predmetov splošnih in poklicnih programov
- **leposlovna** besedila
  - ki so glede na knjižnično izposojo in/ali prodajo bolj brana
  - literatura starejšega izvora, ki pa ima še vedno veliko recepcijo v okviru obveznega šolskega branja
- ta besedila bodo postala integralni del nadgrajenih korpusov

Usmerjeno zbiranje novih gradiv - 2

- besedila izbranih besedilodajalcev z največjo besedilno produkcijo
  - **novičarski portali** (rtvslo.si, 24ur.com, siol.net, žurnal24.si, sta.si itd.)
  - **dnevni časopisi** (delo.si, dnevnik.si, vecer.si itd.)
- podkorpus **GF-novice 2010**
  - samostojna (pod)enota
  - izhodiščno leto objave besedil: 2010
  - obstoječi korpusi
    - slWaC (Erjavec in Ljubešić 2014)
    - newsfeed (IJS, EventRegistry)
    - drugi spletni korpusi
- pogodbena razmerja z besedilodajalci

Strojna obdelava

- orodja, razvita v projektu "Sporazumevanje v slovenskem jeziku"
- XML Text Encoding Initiative P5
- označevanje: Obeliks
  - GitHub: https://github.com/mgrcar/Obeliks

- Novost:
  - tokenizator kot samostojni modul (Miha Grčar)
  - eksperiment z metaoznačevalnikom Obeliks & Nikoliks

Časovni okvir

- 31. december 2016 -> pomlad 2017
  - Kazalnik 4
    - vmesna serija korpusov Gigafida 1.1 in podkorpus GF-novice 2010 1.0 v formatu XML in jezikoslovno označena, dostopna v repozitoriju CLARIN.SI
  - Kazalnik 5
    - korpusa Gigafida 1.1 in Kres 1.1 dostopna v spletnih konkordančnikih
- 31. december 2018
  - Kazalnik 7
    - končna serija korpusov Gigafida 2.0 in podkorpus GF-novice 2010 2.0, v formatu XML in jezikoslovno označena, dostopna v repozitoriju CLARIN.SI
  - Kazalnik 8
    - končna korpusa Gigafida 2.0 in Kres 2.0 dostopna v spletnih konkordančnikih

Vsebinske spremembe

- deduplikacija
  - odstranjevanja dvojnikov na obstoječih besedilih, saj se v besedilih, ki izhajajo iz tiskanih medijev, pojavljajo ponavljajoči se deli besedil, ki v nekaterih primerih izkrivljajo statistične podatke pri poizvedbah po celotnem korpusu
- dodatno tehnično čiščenje
  - besedilo brez presledkov
  - neuspešno prepoznavanje znakov (UTF-8)
  - itd.
- prepoznavanje in izločanje nestandardnih besedil

Standardno in nestandardno

- Tri kategorije:
  - javno objavljena integralna leposlovna in stvarna besedila, revije, časopisi in podobna besedila
  - besedila iz prepoznavnih medijev, ki se iz različnih razlogov odločajo za odmik od standarda
  - računalniško posredovano komunikacijo, ki je značilna za spletne medije – socialna omrežja, forume ipd.

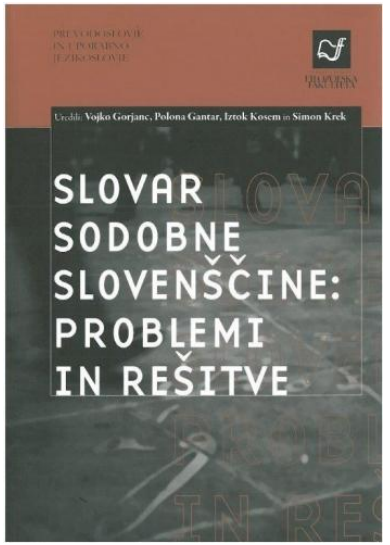
- Gigafida 2.0 = prva kategorija
  - namen nadgradnje je oblikovati korpus standardne slovenščine, kot se definira v sociolingvističnih študijah

Korpus standardne pisne slovenščine

- Gigafida 2.0 (Kres, ccGigafida, ccKres)
  - standardna pisna slovenščina (splošno besedišče)
- Gos (+Gos 2.0)
  - govorjena slovenščina
- slWaC
  - spletna (standardna in nestandardna) slovenščina
- JANES
  - računalniško posredovana (nestandardna) slovenščina
- KAS
  - akademska slovenščina (specializirano besedišče, terminologija)
- drugi (manjši, specializirani itd.) korpusi

Standardna in knjižna slovenščina

- Gorjanc, V., Krek, S. in Popič, D. *Med ideologijo knjižnega in standardnega jezika*. V: Slovar sodobne slovenščine: problemi in rešitve, Filozofska fakulteta 2015.
- Krek, S. *Standardni in knjižni jezik – drugi poskus*. V: Obdobja 34: Slovnica in slovar – aktualni jezikovni opis, Filozofska fakulteta 2015.
- Stabej, M., Dobrovoljc, H., Krek, S., Gantar, P., Popič, D., Arhar Holdt, Š., Fišer, D., Robnik Šikonja, M. *Slovenščina JANES: pogovorna, nestandardna, spletna ali spretna?* V: Slovenščina 2.0, 4 (2).



Po projektu

- skladenjsko razčlenjevanje
- označevanje semantičnih vlog (Semantic Role Labeling)
- prepoznavanje imenskih entitet & wikifikacija (http://wikifier.org/)

- vezljivostni slovar (skladenjski vzorci)
- semantične besedne skice
- luščenje definicij
- itd.