Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres

Simon Krek,[+]* Polona Gantar†, Špela Arhar Holdt[+]†‡, Vojko Gorjanc†

* Laboratorij za umetno inteligenco, Institut »Jožef Stefan«, Jamova 29, 1000 Ljubljana
[+] Center za jezikovne vire in tehnologije, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana
simon.krek@guest.arnes.si
† Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff.uni-ljs.si, vojko.gorjanc@ff.uni-lj.si
‡ Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, 4220 Škofja Loka
spela.arhar@trojina.si

1 Uvod

Prispevek opisuje projekt[1] nadgradnje korpusa Gigafida ter korpusov Kres, ccGigafida in ccKres. Zadnji trije korpusi bodo iz prvega izpeljani po metodologiji, zastavljeni v projektu Sporazumevanje v slovenskem jeziku (SSJ),[2] v okviru katerega so bili izdelani korpusi prve različice. V drugi različici bo korpus nadgrajen predvsem količinsko, nekatere spremembe pa so predvidene tudi na vsebinski ravni, predvsem s konceptualnim razlikovanjem med deli korpusa, ki po avtorski intenci spadajo v jezikovni standard, ter tistimi, ki segajo izven njega. Na tehnični ravni bo korpus v celoti ponovno jezikoslovno procesiran z nadgrajenimi orodji. V prispevku opišemo izhodiščni projektni načrt in specifikacije, izdelane v prvi fazi projekta.

2 Projektni načrt

Zaradi omejenih sredstev je projekt nadgradnje usmerjen predvsem v tiste segmente gradnje korpusov, ki glede na osnovni namen lahko največ prispevajo h končnemu cilju. Korpusi Gigafida, Kres, ccGigafida in ccKres predstavljajo organsko celoto, zato je o njih smiselno razmišljati kot o seriji korpusov, ki ima izvor v korpusu Gigafida, pri čemer so pri izdelavi izvedenih vzorčenih korpusov (Kres, ccGigafida in ccKres) vedno lahko uporabljeni identični mehanizmi kot v primeru izhodiščne izdelave v okviru projekta SSJ. Ena od pomembnih značilnosti korpusov SSJ so urejena pravna razmerja z besedilodajalci, ki omogočajo javni dostop do vsebine in nadaljnjo distribucijo korpusov pod pogodbeno dogovorjenimi pogoji. V projektu nadgradnje bodo podobne možnosti za javno objavo in nadaljnjo distribucijo korpusov ohranjene, zato se projekt osredotoča na selektivno ciljno zbiranje novih gradiv glede na ugotovljene pomanjkljivosti obstoječih korpusov, ne pa na splošno zbiranje vseh gradiv po načelih, ki so bili uporabljeni pri gradnji korpusa Gigafida in njegovih predhodnikov, korpusov FIDA in FidaPLUS. Drugi segment, ki ga omogoča finančni okvir projekta in lahko prispeva k izboljšanju korpusov, je nova oz. dodatna računalniška obdelava že obstoječih in novih gradiv na podlagi analiz, ki so bile opravljene po izdelavi korpusov. Ker je namen projekta čim širša uporaba nadgrajenih korpusov za različne namene, je vanj vključeno tudi omogočanje javnega dostopa do rezultatov in njihova distribucija po modelu iz projekta SSJ. Projekt nadgradnje korpusov Gigafida, Kres, ccGigafida in ccKres ima torej tri cilje: (a) usmerjeno zbiranje novih gradiv; (b) strojna obdelava novih (in obstoječih) gradiv in (c) javna dostopnost nadgrajenih korpusov, distribucija in diseminacija. V poglavju o specifikacijah opišemo prva dva cilja.

3 Specifikacije

2.1 Zbiranje novih gradiv

Pri izpolnjevanju cilja zbiranja novih gradiv delimo besedila na dva tipa: (a) besedila, za katera je bilo glede na tip, zvrst ali druge kriterije po analizi korpusov Gigafida in Kres ugotovljeno, da so podreprezentirana in (b) besedila izbranih spletnih besedilodajalcev z večjo produkcijo (npr. novičarski portali, dnevni časopisi ipd.), ki zagotavljajo večjo aktualnost korpusnega gradiva.

V prvi kategoriji bodo zbirana predvsem šolska gradiva, tj. (prosto dostopni) učbeniki, delovni zvezki ter sorodna učencem ter dijakom namenjena besedila vseh šolskih predmetov splošnih in poklicnih programov (osnovne šole, gimnazije, poklicne srednje šole). Kot druga večja skupina v to kategorijo spadajo tudi

[1] Projekt financira Ministrstvo za kulturo v letih 2015–2018 v okviru pogodbe št. 33400-15-141007 med ministrstvom in Univerzo v Ljubljani. Izvajalec je Center za jezikovne vire in tehnologije Univerze v Ljubljani (http://www.cjvt.si/).

[2] http://www.slovenscina.eu/.

leposlovna besedila, predvsem tista, ki so glede na podatke o knjižnični izposoji in/ali prodajanosti bolj brana. Med njimi je tudi literatura starejšega izvora, ki pa ima še vedno visoko recepcijo v okviru obveznega šolskega branja. Ta besedila bodo postala integralni del nadgrajenih korpusov Gigafida/Kres/ccGigafida/ccKres 2.0.

V drugi kategoriji bodo zbrana besedila izbranih besedilodajalcev z največjo besedilno produkcijo. Izpostavljeni so predvsem novičarski portali (rtvslo.si, 24ur.com, siol.net, žurnal24.si, sta.si) ter dnevni časopisi (delo.si, dnevnik.si, vecer.si itd.). Iz njih bo sestavljen samostojen podkorpus Novice, ki bo vključen v serijo korpusov Gigafida 2.0, vendar bo zaradi omejene žanrske raznovrstnosti besedil ostal samostojna (pod)enota. Izhodiščno leto objave besedil pri tem podkorpusu je 2010.

Skupna ciljna vsota besed v korpusu Gigafida 2.0 (skupaj s podkorpusom Novice) je 1,5 milijarde besed. Število besed v korpusih Kres, ccGigafida in ccKres po metodologiji iz (Logar in dr., 2012) izhaja iz izhodiščne številke: Kres in ccGigafida 2.0 po 150 milijonov besed, ccKres 15 milijonov besed.

2.2 Strojna obdelava novih in obstoječih gradiv

Jezikoslovno označevanje: od časa označevanja besedil v izvirni seriji korpusov (Grčar in dr. 2012) je bil za slovenščino razvit nov statistični označevalnik. Preliminarni testi so pokazali, da bi natančnejše označevanje lahko dosegli z uporabo metaoznačevalnika, ki upošteva odločitve obeh označevalnikov. Z metaoznačevalnikom, katerega izdelava je predvidena v letu 2016, bo ponovno označen celoten korpus.

Deduplikacija: raba korpusa Gigafida po objavi je pokazala, da bi bilo smiselno proces odstranjevanja dvojnikov izvesti tudi 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 (Logar in dr., 2015). Tipičen primer takih besedil so radijski in televizijski programi, ki so z isto vsebino objavljeni v različnih virih. V procesu priprave serije korpusov Gigafida 2.0 bo deduplikacija izvedena tudi na obstoječih besedilih.

2.3 Jezikovni standard

Trenutno Gigafida in Kres prinašata tako besedila, za katere je glede na okoliščine in medij objave mogoče sklepati, da je bil avtorjev namen pisati v standardnem jeziku, kot besedila, pri katerih takšno sklepanje ni mogoče. Ker je namen nadgradnje oblikovati korpus standardne slovenščine, kot se definira v sociolingvističnih študijah (Cooper 1989; v slovenskem prostoru zlasti Krek 2015; Gorjanc et al. 2015), bo v procesu obdelave obstoječih in novih besedil opravljeno segmentiranje korpusnih dokumentov v tri kategorije. V prvo kategorijo spadajo javno objavljena integralna leposlovna in stvarna besedila, revije, časopisi in podobna besedila (znanstvena besedila, zakonodaja itd.). V drugo kategorijo spadajo predvsem besedila iz prepoznavnih medijev, ki se iz različnih razlogov odločajo za odmik od standarda – najbolj značilni predstavnik kategorije je Novi Matajur, ki je zapisan v regionalni varianti slovenščine. V zadnjo kategorijo uvrščamo predvsem računalniško posredovano komunikacijo, ki je značilna za spletne medije – socialna omrežja, forume ipd. Segmentacija bo opravljena ročno glede na izvor besedila, poleg tega bodo besedila preverjena tudi strojno s prepoznavanjem nestandardnih prvin (Ljubešić et al., 2015).

4 Zaključek

Gigafida, Kres, ccGigafida in ccKres so kot referenčni korpusi za slovenski jezik glavni vir za izvedbo jezikoslovnih raziskav oz. pripravo uporabnojezikoslovnih in jezikovnotehnoloških izdelkov. Na drugi strani sta Gigafida in Kres v vmesnikih SSJ priljubljeno in pogosto uporabljano orodje tudi v širši javnosti, npr. med prevajalci, lektorji, učitelji. Specifični status in vloga teh virov v prostoru zahtevata njihovo kontinuirano nadgrajevanje in opisani projekt odgovarja na nekatere od glavnih identificiranih potreb. S tem predstavlja težko pričakovani korak naprej – čeprav seznam nalog in želja za nadaljnji razvoj ostaja dolg in raznolik.

Literatura

Cooper, Robert L. 1989. Language Planning and Social Change. Cambridge: Cambridge University Press.

Erjavec, Tomaž in Darja Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. *Družbena funkcijskost jezika: (vidiki, merila, opredelitve), Obdobja 32*. Ljubljana: Znanstvena založba Filozofske fakultete, 109–116.

Gorjanc, Vojko, Simon Krek in Damjan Popič. 2015. *Med ideologijo knjižnega in standardnega jezika*. Ljubljana: Znanstvena založba Filozofske fakultete. 32–48.

Krek, Simon. 2015. Standardni in knjižni jezik – drugi poskus. Smolej, M. (ur.). *Obdobja 34: Slovnica in slovar – aktualni jezikovni opis*. Ljubljana: Znanstvena založba Filozofske fakultete, 401–407.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. *10th International Conference on Recent Advances in Natural Language Processing*: *Proceedings of RANLP 2015 Conference*, 7–9 September 2015, Hissar, Bulgaria. Hissar: 371–378.

Logar Berginc Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Nataša Logar, Kaja Dobrovoljc in Špela Arhar Holdt. 2015. Gigafida: interpretacija korpusnih podatkov. V: M. Smolej, ur., *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis*, 2. del, str. 467-477. Ljubljana: Znanstvena založba Filozofske fakultete.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012*. Ljubljana: Institut Jožef Stefan.