KORPUSNOJEZIKOSLOVNI MO(NU)MENTI: *KORPUSI SLOVENSKEGA JEZIKA GIGAFIDA, KRES, ccGIGAFIDA IN ccKRES: GRADNJA, VSEBINA, UPORABA*

Damjan POPIČ
Univerza v Ljubljani, Filozofska fakulteta

Popič, D. (2013): Korpusnojezikoslovni mo(nu)menti: Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba. Slovenščina 2.0, 1 (1): 176–180.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_09.pdf.

1 SUBJEKTIVNI DEL

»Pripravi naj se par strani dolgo besedilo o namenu korpusa in razlogih zbiranja materialov – namenjeno ljudem in institucijam, od katerih bomo skušali dobiti material.«

Takšen je bil pred več kot 15 leti sklep štiričlanske komisije, ki si je naložila nalogo, da sestavi referenčni korpus slovenskega jezika. Načrt je bil visokoleteč, tipično zaznamovan z začetnimi težavami (iskanjem darovalcev besedil, izdelavo predlog ipd.), v veliki meri takšnimi, s kakršnimi se danes slovenskim korpusnim raziskovalcem prav zaradi tega začetnega truda ni treba več (toliko) ubadati.

Po dolgi poti od Fide, ki je služila kot nekakšen testni poligon za vpeljavo korpusnojezikoslovnih praks v slovenski prostor in sprotno odločanje o posameznih problemih, prek Nove besede z nekoliko drugačnimi (stremljivejšimi?) izhodišči, do FidePLUS in Gigafide: po dobrih 15 letih je slovenska korpusnojezikoslovna krajina izrazito bogatejša, monografija *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba* pa je pripraven, v prijetno obsežni meri pa tudi objektiven monument nedavno dodanim razsežnostim slovenskega korpusno podprtega in korpusno usmerjenega jezikoslovnega raziskovanja.

Za začetne misli v letu 1997 bi le stežka trdili, da so bile revolucionarne v svetovnem merilu, za slovensko okolje pa so bile seveda izjemnega pomena. V tem oziru je slovensko korpusno jezikoslovje v relativno kratkem času zelo napredovalo, saj ni dobilo le številnih besedilnih korpusov za slovenski jezik, temveč tudi odprte jezikovne tehnologije, ki na tej infrastrukturi temeljijo.

Vsemu temu navkljub pa se je težko otresti občutka, da je slovensko korpusno jezikoslovje v slovenski družbi doseglo (pre)malo, in sicer zaradi mesta, ki ga ima v slovenistični jezikoslovni misli. Ob vseh jezikovnih tehnologijah, trudu in rezultatih je računalniško jezikoslovje še vedno zgolj – če sploh – sekundant jezikovnemu občutku, ki slovenistično jezikoslovje nezmotljivo vodi tudi v tretjem tisočletju. Ali je za to kriva jezikovnotehnološka stroka sama ali pa je zgolj žrtev specifik slovenske situacije – v kateri je denimo mogoče besedo zakonodajno novelirati v slovarju na podlagi lepotnega tekmovanja z izpitno komisijo, če le imate besedotvorni talent –, je težko reči s kakršno koli objektivno veljavo, zagotovo pa drži, da najplemenitejši in zdaleč najpomembnejši cilj korpusnega jezikoslovja – opolnomočenje jezikovnega uporabnika ter razlaščanje jezikoslovnega patricija – v slovenskem okolju še ni bil izpolnjen.

Tako je tudi v letu 2013 slovenistična stroka marsikje izrazito tradicionalistična, četudi se na prvi pogled ne zdi tako in četudi se je v slovenskem prostoru proti temu sistematično borilo vse od šestdesetih let prejšnjega stoletja, jezikovne tehnologije pa – če so že razumljene kot relevanten del slovenističnega raziskovanja – sodijo daleč na njeno obrobje. Medtem ožja slovenistična *de facto* normodajna stroka na področna in širša družbena vprašanja glede rabe slovenskega jezika še vedno odgovarja z zavidljivo samozavestjo ter bolestnim ljubosumjem, ki ga poji prepričanje, da je edina primerna za normiranje vsega, česar se spomni slovenščina. Pri tem se ta smer tudi hermetično zapira, novosti – zlasti tehnološke – pa so razumljene kot grožnje njeni eksistenci (ob bizarnem nerazumevanju paradoksa, da prav izolacija ogroža eksistenco digitalnih ksenofobov, v končni fazi pa najbrž tudi domorodcev digitalnega sveta). S tem je celotna stroka seveda relativizirana in potisnjena na obrobje družbene relevance, slovenščina

kot komunikacijsko orodje pa je v nevarnosti, da v prihodnosti opeša pri odgovarjanju na izzive informacijske družbe.

Polariziranost širše slovenistične stroke meji na dvostrankarski sistem z doslednim sledenjem strankini liniji. Monografija *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba* prinaša podroben vpogled v nove jezikovne tehnologije, vendar pa je za slovenistično spravo pomembnejša njena snovna podstat – brezkompromisna ideologija o odprti družbi in dostopnosti informacij za neprivilegirane.

2 OBJEKTIVNI DEL

Monografija prinaša podroben vpogled v drobovje in gradnjo Gigafide ter treh podkorpusov v okviru projekta Sporazumevanje v slovenskem jeziku, potem ko je javno predstavitev že doživel korpus Šolar. Zelo podrobno sta predstavljena proces zbiranja in tipologija besedil v korpusu Gigafida, veliko prostora je namenjenega tudi konkordančniku in zapisu.

Publikacija zelo podrobno predstavi gradnjo in sestavo korpusov v številkah in opisih delovnih tokov, a bržkone so najpomembnejše informacije o odločevalskih procesih, ki neizogibno spremljajo vsako korpusno gradnjo ter imajo kljub začrtani objektivnosti močan vpliv na poznejše dogajanje in rezultate – v slovenskem okolju je to na primer dobro vidno, če primerjamo strukturo in poslanstvo korpusov FidaPLUS ter Nova beseda.

Celostna dikcija monografije je še toliko bolj dobrodošla zaradi tega, ker je predstavljeno širše ozadje slovenske korpusnojezikoslovne zgodbe, vse od Fide naprej, in ker je marsikje opremljena z iskrenimi komentarji o posameznih odločitvah pri konkretnih problemih. Prav ta pozitivna plat monografije pa je morda tudi njena pomanjkljivost – ker si prizadeva podati čim več informacij, se zdi kot slovenski korpusnojezikoslovni omnibus, ki lahko na trenutke pozabi, koga nagovarja. Tako utegne manj izkušenega bralca, ki bo v knjigi iskal nazivno *uporabo*, hitro pregnati, obenem pa poglavje o Fidi deluje nekoliko antološko in nerelevantno glede na (predvideni) namen knjige – morda bi bilo glede na njun pomen bolj na mestu razširjeno poglavje o prosto

dostopnih zbirkah ccGigafida in ccKRES.

Kljub temu pa gre za zelo relevantno delo, pravzaprav kar »uporabniško dokumentacijo« in monument sodobnega slovenskega korpusnega jezikoslovja. Za slednje sta korpusa Gigafida in KRES izjemen dosežek, a ker v ciljnem okolju jezikovne tehnologije v resnici prednjačijo pred progresivnostjo celostne jezikoslovne misli, optimizem zbujata predvsem prosti zbirki ccGigafida in ccKRES (kratica cc zaznamuje licenco Creative Commons, ki velja za zbirko besedil v formatu XML in jo lahko prosto prenesemo s spleta, seveda ob upoštevanju – zelo liberalnih – določil te licence).

Prosti zbirki sta namenjeni raziskovalcem, tako pri nas kot na tujem, ki bi želeli besedilne zbirke v obliki podatkovne baze uporabiti za razvijanje nadaljnjih jezikovnih tehnologij, in tistim, ki pri svojih korpusnojezikoslovnih analizah trčijo ob meje zmogljivosti konkordančnika korpusov Gigafida in KRES. Obe prosti zbirki vsebujeta okoli 9 odstotkov besedil svojih matičnih korpusov, razlog za to razmerje pa je v omejitvi prenosa avtorskih pravic s strani besedilodajalcev. Odločitev, da se zbrano gradivo ponudi v uporabo raziskovalcem, kaže na nekoliko atipično miselnost za ožje ciljno okolje, prav to pa je ideologija, ki spodbuja sodelovanje in odprto informacijsko družbo, še več: povsem odprto in transparentno razpolaganje z vso intelektualno lastnino, ki nastane z javnofinančno podporo ter javnosti posledično tudi pripada.

Poleg vnašanja svežih ideologij in miselnosti v slovenski raziskovalni prostor pa gre največji plus publikacije pripisati težnji po brezpogojni objektivnosti, ki je bila monografiji položena v zibko. Le s tovrstnimi praksami je mogoče vplivati na ugled celotne slovenistike kot resne znanstvene panoge kot tudi na mesto računalniškega jezikoslovja v slovenskem jezikoslovju.

To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva-Deljenje pod enakimi pogoji 2.5 Slovenija.

This work is licensed under the Creative Commons Attribution ShareAlike 2.5 License Slovenia.

http://creativecommons.org/licenses/by-sa/2.5/si/