SLOVENSKO PRAVOPISJE IN KORPUSI

Peter Weiss
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU, Ljubljana

UDK 811.163.6'35:811.163.6'271.1:81'322

Slovensko korpusno jezikoslovje je v zadnjih dveh desetletjih pri obravnavi pisane leksike in pregledu nad njo gradivsko res zelo napredovalo, vendar pa je ob tem ostal zelo zanemarjen pogled na slovenski pravopis. Zdajšnji korpusi izkazujejo nekatera pravopisne pojave glede na izvirnike v okrnjeni obliki (nenatančno povzeta ločila, zveze presledkov in ločil, simboli) in na ravni, ki je uporabnikom s trenutnimi računalniškimi iskalnimi orodji nedosegljiva (nizi ločil, simboli).

slovenski pravopis, korpusno jezikoslovje, pisani jezik, govorjeni jezik

In terms of material, Slovene corpus linguistics has taken huge strides forward in the last two decades in its awareness and treatment of written lexica, but at the same time Slovene orthography has been considerably neglected. Current corpora show certain orthographical phenomena in relation to the original versions in truncated form (imprecise punctuation, relations between spaces and punctuations marks, symbols) and at a level that is inaccessible to users with the current computer search tools (strings of punctuation marks, symbols).

Slovene orthography, corpus linguistics, written language, spoken language

1

Pravopis kakega jezika obravnava v ožjem smislu (1) predpisano zaporedje črk oz. izbiro črk, (2) pisanje skupaj oz. narazen, (3) pisanje z veliko in malo začetnico in (4) ločila (Christian Stetter po Levin-Steinmann 2007: 41). V slovenskih razmerah je slovarski del pravopisnega priročnika vsaj po drugi svetovni vojni (z izdajama v letih 1950 in 1962 s ponatisi) nadomeščal slovar knjižnega jezika, ki ga nismo imeli in ki je začel kot *Slovar slovenskega knjižnega jezika* (SSKJ) izhajati šele leta 1970, končan pa je bil dobri dve desetletji pozneje. Zato je bil pravopisni slovar razširjen s pravorečnimi, oblikoslovnimi, stilističnimi in sopomenskimi podatki, čeprav so bili slednji zapisani bolj kot skrajnosti *pravilno – nepravilno* na premici s prav malo odtenki. Zadnja izdaja *Slovenskega pravopisa* (iz leta 2001) kljub medtem končanemu *Slovarju slovenskega knjižnega jezika* nadaljuje širitev na stilistično področje ob ohranitvi vsega drugega, v marsičem z naslonitvijo na SSKJ. V slovenskih razmerah tak pravopis ni več specializirani jezikovni priročnik, ampak poskuša biti osrednje priročniško orodje za jezikovno načrtovanje. Ker je to preobloženo z najraznovrstnejšimi podatki, je komajda še obvladljivo za uporabnike in pisce; prevelik obseg zahteva kleščenje vsaj pri prikazovanju pomenov in pri označevanju besedja.

2

Odnos uporabnikov slovenščine do pravopisa knjižnega jezika je protisloven, kar ni nikakršna posebnost zadnjega časa in tudi ne samo naših krajev. Pravopisna določila, ki so utemeljena in zapisana, npr. v *Slovenskem pravopisu* kot zbirki objavljenih pravil s konkretizacijo v pravopisnem slovarju, so pri uporabnikih

slovenskega jezika večinoma upoštevana: določila rabo usmerjajo, hkrati pa jih raba v jedru potrjuje. Sožitje obojega je jezikovna norma. V posameznostih in v vsakdanji praksi pa so določila marsikdaj nezadostna. Če uporabnik ustrezne rešitve ne najde v pravopisnih pravilih in priloženem slovarju, se (lahko) ravna po vzorcih, da bi se približal tistemu, kar sorodnega je v pravilih in slovarju že zajeto. Na ta način se iščejo nove, boljše rešitve, kar je dobro, saj pri slabo premišljenem (SP 2003 predpisuje *okolijski* namesto široko rabljenega in brez težav izgovorljivega *okoljski*, pač v skladu z edino možnim *ladijski*), nesistemskem (v SP 2003 *obsoteljski* nam. *obsotelski*, kar je v nasprotju s pravopisno uveljavljenim *podeželski*; vezalne zloženke, ki naj bi se po novem pisale z vezajem, npr. *kupo-prodaja* in *kupo-prodajen*, toda *gluhonem* in *gluhonemnica*, *severozahod*, *knezoškof*, *generalmajor* ...; o vsem tem Weiss 2003: 203–204) in napačnem (*daljnjovzhoden* v slovarskem delu SP 2001, kar je bilo v izdaji SP 2003 popravljeno v *daljnovzhoden*) v pravopisu pač ni mogoče vztrajati. Vendar pa nekateri uporabniki slovenskega jezika v posameznostih pravopisno normo knjižne slovenščine bolj ali manj zavestno zavračajo in pišejo v skladu s svojimi pogledi na jezikovno predpisovanje ali pa katerega od določil ne sprejemajo, recimo zaradi odločitve za sleng (npr. v esemesih), zaradi pretirane vneme po jezikovni pravilnosti (v dnevniku *Delo* uporabljano *talib* nam. *taliban*, kar da je množinska oblika že v jeziku dajalcu; prim. Weiss 2002) ali preprosto zaradi odpora proti normiranemu, ki v leposlovnih besedilih s stilizacijami dosega umetniške učinke (ali pa deluje mimo njih). Če jezikovno svobodo jemljemo resno, jo je treba dopustiti vsakomur, hkrati pa ne smemo nikomur kratiti pravice do jezikovne solidarnosti. Ta se izraža v skupni rabi jezika in privrženosti normi (ob hkratni možnosti do utemeljenega odstopanja od nje), ki bi ju resno jezikovno načrtovanje moralo spodbujati.

3

Jezikovni korpus kot sistematična zbirka besedil, ki dokumentira rabo jezika ali jezikovnih različkov, je lahko le namenski in zamejen in ne more dokumentirati jezika v celoti, saj je njegovih pojavnih oblik za hkratni zajem preprosto preveč (Bergenholtz-Mugdan 1989: predvsem 142–143; 1990: predvsem 1620–1621).

Slovensko korpusno jezikoslovje je sploh v zadnjem desetletju z računalniško podporo pri obravnavi zapisanih besedil, besedja iz njih in pregledu nad njim ter nad slovničnimi, vezljivostnimi ipd. podatki o upoštevanih besedah in besednih zvezah gradivsko zelo napredovalo (prim. Gorjanc 2005), vendar pa je ob tem ostalo tako rekoč povsem zanemarjeno korpusno upoštevanje ločil in simbolov, kar vodi v pomanjkljive opise rabe ločil, ki bi jih morali najti v pravopisnih priročnikih. Ločila kot opazovani predmet v korpusih je težko proučevati, ker so ločila kljub (možni) standardizaciji (v računalnikih in v elektronskih besedilih) glede na izvirna besedila poenostavljena, korpusni iskalniki pa ne omogočajo iskanja po njih. Taki so npr. začetni in končni narekovaji (» «, „ “ in “ ”), ki po navadi sovpadejo (v ⇓), pomišljaj (‑) ali dolgi pomišljaj (—) in vezaj (-), ki v korpusih prav tako sovpadejo (v vezaj), tri pike (...), ki se v računalniških programih lahko pišejo tudi kot en sam znak (v SP 2003 imenovan »tripičje«), standardiziran v unikodu, znak za ‘(kotna) stopinja’ ali ‘stopinja Celzija’ (°, namesto katerega nekateri pišejo tudi nadpisani o, tj. ᵒ) in opuščaji (’, ‘ in ʻ), ki jih v korpusih nadomešča resica (⁇). V pravopisne priročnike zato ne bo mogoče zapisati najnovejšega korpusnega stanja ločil in ponazarjalnih primerov, torej jih bo treba še naprej zbirati namensko ali jih izpisovati ročno in torej izbirno, kar je v primerjavi s korpusnim pristopom velika slabost. (Po ločilih, kakršna so pač zapisana v korpusu Nova beseda, bo s prilagoditvijo korpusa in iskalnika menda vendarle mogoče iskati, kot je

oktobra 2009 zagotovil njegov skrbnik Primož Jakopin.) Pravopis naj bi opisal rabo ločil v kombinacijah, recimo v nizu vprašaj in končni narekovaj (in morda še vejica, ki jo nekateri pišejo, drugi pa ne) na koncu dobesednega navedka premega govora, uvedenega z dvopičjem, pred novim stavkom v isti povedi: *Kot bi hotel zavpiti: »Mi pa že nismo takšni!«, a bi potem izdahnil le potihoma: »Mi pa že nismo takšni?«* (*Delo – Sobotna priloga* 19. 7. 2003. 27), pa tudi: *Palček in veverica sta kričala: »Nehaj, nehaj!« velikan pa ju sploh ni slišal*. (*Pikapolonica: revija za radovedne, ustvarjalne in igrive otroke* 4 (2002–2003), št. 11–12. 41). Zdajšnja največja slovenska korpusa – FidaPLUS (http://www. fidaplus.net/) in Nova beseda (http://bos.zrc-sazu.si/s_beseda.html) – poleg črkovnih in številskih znakov sicer izkazujeta tudi ločila, ki se jih posamezne v korpusu FidaPLUS da iskati. Vendar pa se v korpusih ne da najti nizov ločil (npr. *?!«,*) ali nekaterih problematično zapisanih ločil (npr. - namesto –) ali zveze presledka, ki naj bi bil kot pravopisno marsikdaj problematičen vreden upoštevanja tudi v zbirkah besedil, in ločila (npr. <presledek>...), kar vse z nadpisanimi ali podpisanimi črkovno-številskimi znaki (npr. v m^2, CO_2, 19^h, 8^{15}) vred ostaja v korpusih na ravni, ki je jezikoslovcem in drugim uporabnikom s trenutnimi računalniškimi iskalnimi orodji nedosegljiva v obliki, ki bi izkazovala tisto v izvirniku.

Za (slovensko) pravopisje bo ob tem potrebno tudi širše, usmerjeno in prilagojeno zajemanje ustreznih pisnih pojavov v slovenskih besedilih, kot je npr. členjenje besedil na naslovje in odstavke, kakršni so v izvirnikih, kar bi pomagalo npr. pri raziskovanju ločil pri naštevalnih enotah.

Največji primanjkljaj korpusnega pristopa pri zajemanju virov in primerov za slovensko pravopisno in tudi siceršnjo slovaropisno rabo se kaže pri težavnih glasoslovnih mestih (kot sta npr. posebni izgovor črke l, ki mora biti v priročniku, kot je pravopis, jasno razviden, ali naglasno mesto *vôdami* ali *vodámi*) in pri stilistični predstavitvi besedja in oblik (z najrazličnejšimi označevalniki). Korpus govorjenih besedil je za slovenščino trenutno še želja. Poleg tega naši nastajajoči ali zasnovani korpusi očitno ne bodo omogočali glasoslovne analize govora. Stabej in Vitez (2000: 81) sta napovedala, da bo korpus KGB omogočal morfološko, skladenjsko in semantično analizo govora, pa tudi iz objav Jane Zemljarič Miklavčič (2007) in Darinke Verdonik (2007) ter iz dokumenta Sporazumevanje v slovenskem jeziku – govorni korpus (http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K4.aspx, dostop 15. 10. 2009) ne izhaja za glasoslovje (ki je v slovenskem pravopisu vse pravorečno, torej poleg posebnega izgovora posameznih črk tudi vse naglasno, recimo z razlikovanjem med *meščán* in *meščàn*) nič obetavnega.

Še naprej bodo korpusi v slovaropisju (in slovarski del pravopisa, kakršen je zadnji, iz leta 2003, ni nikakršna izjema) le omejeno uporabni za označevanje besedja, torej se bodo slovaropisci morali večinoma zanašati na opazovanje drugih in sebe, vrednost označevalnikov v slovarjih pa je tako ali tako le relativna, ne absolutna (Franz Josef Hausmann v Weiss 2000: 31).

4

Tule ob korpusu spet govorim (prim. Weiss 2001) o podrobnostih (Gorjanc 2005: 54: idejni osnutek Petra Weissa se namreč veliko bolj »ukvarja s tehničnimi detajli, ki pa za samo vsebinsko zasnovo korpusa niso zanimivi«), ki pa vendarle kažejo na relativnost trditev, kako recimo skuša korpus FIDA »posredovati vsestranske informacije o sodobnem slovenskem jeziku, torej z besedili skuša zajeti čim bolj celovito podobo današnje slovenščine, ob jasnem zavedanju, da je nemogoče predvideti in v korpus zajeti vse jezikovne variante« (Gorjanc 2005: 47). Slovaropisci vejo, da se kakovost slovarja ne meri toliko po številu gesel kot po načinu in predstavitvi podatkov v njem, in tudi

pri korpusu niso toliko odločilne stotine milijonov besednih oblik, kar se v zvezi s korpusi pretirano poudarja, kot možnost pridobitve včasih na videz drobnih podatkov, ki v njem zagotovo so. V korpusu je jezik, če nanj res gledamo vsestransko, zajet le delno. Zavedanje tega dejstva se dotika družbene odgovornosti korpusnega jezikoslovja ter odgovornosti korpusnih jezikoslovcev in graditeljev korpusov za svoje delo in navsezadnje tudi za delo drugih (Wiegand, Jesenšek 2006: predvsem 362).

5

Brez korpusov in dosežkov korpusnega jezikoslovja v slovnici, slovaropisju in pravopisju ne gre več, vendar pa bo treba povečati in spodbuditi sodelovanje med graditelji korpusov in korpusnimi jezikoslovci ter slovničarji, slovaropisci in pravopisci. Slednji se kljub upoštevanju korpusnih podatkov v svojih izdelkih ne smejo zbati uporabiti tudi podatkov, pridobljenih na druge načine, če v korpusih, kot vidimo, niso dosegljivi ali pa nikoli ne bodo zadovoljivi (npr. razmerje med vezajem in pomišljajem, ki je v korpusih trenutno zabrisano zaradi prepisa v vezaj v vseh položajih). Posebne, namenske zbirke besedil s kar se da natančnim zajemom pravopisno veljavnih podatkov za raziskave slovenskega pravopisja so ob problemskih izpisih in introspektivnem pristopu možnost, ki bi z izboljšanimi računalniškimi iskalniki dala optimalne rezultate. Korpusi govorjenih besedil bodo uporabni za sestavljanje pravopisnih priročnikov, če bodo dovolj natančno upoštevali ustrezno glasoslovno problematiko. Pri vsem tem (tudi svarilni) izsledki klasičnega korpusnega jezikoslovja seveda lahko samo koristijo. Čakanje na boljše čase na trenutni korpusni strani bi lahko pri raziskovalcih pravopisja, ki potrebujejo podatke o jezikovnih dejstvih za konkretne raziskave v tem trenutku, pomenilo zastoj, ki pa si ga zaradi vendarle skupne, jezikoslovne stroke ne smemo ne želeti ne privoščiti.

Literatura

BERGENHOLTZ, Henning, MUGDAN, Joachim, 1989: Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität. Bátori, István S., Lenders, Winfried, Putschke, Wolfgang (ur.): *Computational Linguistics/ Computerlinguistik* (Handbücher zur Sprach- und Kommunikationswissenschaft 4). Berlin, New York: Walter de Gruyter. 141–149.

BERGENHOLTZ, Henning, MUGDAN, Joachim, 1990: Formen und Probleme der Datenerhebung II: Gegenwartsbezogene synchronische Wörterbücher. Hausmann, Franz Josef, Reichmann, Oskar, Wiegand, Herbert Ernst, Zgusta, Ladislav (ur.): *Wörterbücher/Dictionaries/Dictionnaires* 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 5.2). New York: Walter de Gruyter. 1611–1625.

GORJANC, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

LEVIN-STEINMANN, Anke, 2007: Orthographie und Phraseologie. Burger, Harald, Dobrovol'skij, Dmitrij, Kühn, Peter, Norrick, Neal R. (ur.): *Phraseologie/Phraseology* (Handbücher zur Sprach- und Kommunikationswissenschaft 28.1). Berlin, New York: Walter de Gruyter. 36–41.

STABEJ, Marko, VITEZ, Primož, 2000: KGB (korpus govorjenih besedil) v slovenščini. Bavec, Cene idr. (ur.): *Informacijska družba IS 2000*. Ljubljana: Institut Jožef Stefan. 79–81.

VERDONIK, Darinka, 2007: *Jezikovni elementi spontanosti v pogovoru: diskurzni označevalci in popravljanja*. Maribor: Slavistično društvo.

WEISS, Peter, 2000: Označevanje v slovenskih narečnih slovarjih. Keber, Janez (ur.): *Jezikoslovni zapiski* 6. Ljubljana: Inštitut za slovenski jezik Frana Ramovša. 27–44.

WEISS, Peter, 2001: Slovenski nacionalni korpus Maks na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU: utemeljitev. Keber, Janez (ur.): *Jezikoslovni zapiski* 7/1–2. Ljubljana: Inštitut za slovenski jezik Frana Ramovša. 419–428.

WEISS, Peter, 2002: Talibi in obzorje. *Delo*. *Sobotna priloga* 44/9. 31.

WEISS, Peter, 2003: Slovenski pravopis 2003 – priročnik na stranpoteh slovenskega jezika. Jesenšek, Marko (ur.): *Perspektive slovenistike ob vključevanju v Evropsko zvezo*. Ljubljana: Slavistično društvo Slovenije. 201–206.

WIEGAND, Herbert Ernst, JESENŠEK, Vida, 2006: O družbeni odgovornosti znanstvene leksikografije. Jesenšek, Marko, Zorko, Zinka (ur.): *Jezikovna predanost: akademiku prof. dr. Jožetu Toporišiču ob 80-letnici*. Maribor: Slavistično društvo, Ljubljana: SAZU. 361–378.

ZEMLJARIČ MIKLAVČIČ, Jana, 2007: *Načela oblikovanja govornega korpusa slovenščine*. Doktorska disertacija. Ljubljana.