O korpusnem jezikoslovju

Študije o korpusnem jezikoslovju, Zbornik, Knjižna zbirka Krt 130, ur. V. Gorjanc in S. Krek, Ljubljana, Krtina, 2005, 199 str.

Andreja Žele

IZVLEČEK: Zbornik študij o korpusnem jezikoslovju seznanja z nekaj temeljnimi in hkrati relevantnejšimi študijami o besedilnih korpusih in seveda o korpusnem jezikoslovju, ki je ob gradivu za jezikoslovno analizo izoblikovalo tudi nekaj metodologij, ki skušajo zajeti in predstaviti jezikovno realnost – metodologijo gradnje korpusov, metodologijo korpusne analize in metodologijo novih jezikovnih opisov.

On Corpus Linguistics

ABSTRACT: The studies published in Študije o korpusnem jezikoslovju (eds. V. Gorjanc and. S. Krek, Ljubljana: Krtina, 2005) are some of the most relevant studies discussing text corpora and corpus linguistics. In dealing with materials for linguistic analysis several methodologies were developed in the field of corpus linguistics with the intent to comprise and describe the linguistic reality, e.g. the methodology of corpus building, the methodology of corpus analysis and the methodology of new linguistic descriptions.

0 Zbornik *Študije o korpusnem jezikoslovju* obsega sedem razprav, vsebinsko temeljnih in glede na izbor nekaterih avtorjev (npr. R. Quirk, J. Sinclair) vsaj deloma tudi pionirskih za področje korpusnega jezikoslovja. Z vsebinsko-informativnega vidika je vseh sedem prispevkov (šest prevedenih v slovenščino in en slovenskega avtorja) pretehtano izbranih – predstavljajo namreč dosedanjo kronologijo nastajanja in razvijanja korpusnega jezikoslovja oz. prikazujejo dosedanji razvoj tega jezikoslovnega področja in hkrati nakazujejo smeri nadaljnjega razvoja: *Prispevek k opisu rabe angleškega jezika* (7–27, Radolph Quirk), *Stanje stvari v korpusnem jezikoslovju* (29–57, Geoffrey Leech), *Jezik kot sistem in jezik kot primer*: *korpus kot teoretični konstrukt* (59–79, M.A.K. Halliday), *Prazno besedišče* (81–102, John Sinclair), *Korpusno jezikoslovje in leksikografija* (103–136, Wolfgang Teubert), *Jezikoslovni korpus*: *sredstvo in vir spoznanj* (137–171, František Čermák) in za konec še prispevek slovenskega jezikoslovca V. Gorjanca *V mavrici jezikovnih podatkov* (173–194), ki uvaja tudi v slovenske razmere korpusnega jezikoslovja in je v vsakem primeru uvodni oz. uvajalni, čeprav verjetno po maniri upoštevanja ve'i' ih tujih jezikoslovnih eminenc razvrščen na konec, ponujene bibliografije pri pos meznih

prispevkih pa ponujajo veliko strokovnih virov o korpusnem jezikoslovju ali z njim kakor koli po-/na-vezanih.

1 Temeljnejši poudarki in smernice

– Korpusno jezikoslovje je v slovenskem prostoru z zaključenimi projekti oblikovanja korpusov uspešno končalo prvo in nujno potrebno fazo za nadaljnji razvoj (Gorjanc, 186).

– Besedilni korpusi so neprecenljivi referenčni vir pri vseh vprašanjih, kjer odpovedujeta formalna slovnica in intuicija (Čermák, 153). Korpusno jezikoslovje opazuje jezik kot družbeni pojav in program korpusnega jezikoslovja ni v protislovju s klasičnim jezikoslovjem, temveč samo sebe razume kot njegovo dopolnilo (Teubert, 110, 131).

– Zelo bistvena temeljna predpostavka korpusnega jezikoslovja je v tem, da pomen elementov in segmentov besedila lahko iščemo samo v diskurzu in nikjer drugje (Teubert, 117). Upoštevati je potrebno namreč dejstvo, da »se bo vrojena dinamična spremenljivost jezika v nedogled upirala statičnemu opisnemu aparatu« (Sinclair, 101).

– Osnovno delo z besedilnim korpusom se lahko prikaže v petih stopnjah: 1) identifikacija oblik v besedilu, 2) ugotovitev distribucije oblik in njenih kombinacij z namenom odkriti skladenjske in pomenske enote in njihove kombinacije, vključno s stalnimi, 3) ugotovitev, kako te pomenske enote in njihove kombinacije tvorijo višje pomenske celote in zgradbe, 4) ugotovitev, kako se te višje zgradbe kombinirajo v osnovni besedilni enoti, 5) ugotovitev, kako se določeni izsledki odražajo v zgradbah drugega jezika (Čermák, 155).

– Predvsem v leksikografiji je korpusno jezikoslovje uvedlo nov način dela in tudi razširitev predmetnega področja; prvi primer korpusnega slovarja je Sinclairjev slovar splošnega jezika na osnovi korpusa Cobuild (Teubert, 106).

– V smislu jezikovnosamoumevnega povezovanja slovnice in slovarja R. Quirk (9) med drugim ugotavlja, da »/n/ekaj najbolj plodnih razmišljanj jezikoslovcev v zadnjih letih je bilo na temo medsebojnega prežemanja besedišča in slovnice ter stopnje, do katere sta tako tvorba kot interpretacija fraznih struktur odvisni od neločljive celote pomenskih in slovničnih analogij«.

– Medtem ko strojna oprema skokovito napreduje, ji tehnologija programske opreme prepočasi sledi, še večji problem pa je počasno in zapleteno pravno urejanje avtorskih pravic (Leech, 34).

– Zbirke računalniško berljivih besedilnih zbirk so v tridesetih letih narasle z enega milijona na skoraj tisoč milijonov besed, zato lahko do leta 2021 pričakujemo sorazmerni tisočkratni porast na bilijon besed (Leech, 32).

2 Terminologija in opredelitve

Vsi prispevki so za slovenščino tudi terminološko in opredelitveno relevantni – poleg jezikoslovnih opredelitev so za slovensko jezikoslovje relevantne tudi slovenske terminološke ustreznice. Vsako novo področje prinaša tudi novo terminologijo z novimi opredelitvami, zato je ta zbornik med drugim tudi eden izmed prispevkov k širitvi slovenske jezikoslovne terminologije. V nadaljevanju povzemam nekaj osnov-

nega korpusnega izrazja z opredelitvami – vključno z nekaj osnovnimi opredelitvami bo korpusno izrazje označeno s poševnim tiskom:

– *Referenčni korpus* je osrednji tip korpusa, ki predstavlja določen jezik v čim širšem obsegu njegove pojavnosti in je vezan tudi na določitev parametrov za uravnoteženost v korpusu zajetih besedil na eni strani ter njihovo jezikovno označenostjo v korpusu na drugi (182).

– Korpusi se delijo glede na jezik (število gre v desetine in en jezik ima lahko tudi več korpusov), besedilne tipe (npr. splošni/nespecifični in specializirani korpusi /sinhroni – diahroni, terminološki, narečni/), glede na vrsto prenosnika (korpusi pisnega ali govorjenega jezika); našteta delitvena merila določajo t. i. *podkorpuse* (142, 143).

– *Elektronsko knjižnico* oz. *tekstoteko* kot prosto zbirko besedil je potrebno razlikovati od pravega korpusa z različnimi shranjevalnimi tipologijami in z različnimi stopnjami oblikovne in skladenjske označenosti (141).

– *Podatkovna zbirka* je oblika obdelovanja in urejanja korpusnih podatkov glede na različne potrebe. Je navadno relacijskega tipa z uporabo individualno izdelanih ali razširjenih komercialno uspešnih programov podatkovnih zbirk (150).

– *Lematizator* je program, ki sam ali v povezavi z drugim programom, npr. polnobesedilno podatkovno zbirko, zmore vse besedne oblike besede zbrati pod skupno lemo, npr. pod imenovalnik ali nedoločnik (150).

– »*Prazno*« *besedišče* (Sinclair, 99) je besedišče živega jezika, ki se uči iz besedil, sega torej prek intuicije posameznika, in se nenehno posodablja – njegovo nasprotje je klasično pojmovano besedišče kot razširjena baza terminov.

– *Podjezik* je jezik določene skupnosti z zamejeno specializirano rabo na določeno vsebinskospecializirano področje, ki vključuje podvrsto splošnega jezika (Sinclair, 99).

– *Slovnicoslovje* (ang. grammatics, Halliday, 59) proučuje slovnico.

3 Poleg vsebinsko-terminološke informativnosti je zbornik tudi pokazatelj izrazijskih (prevajalskih) zmožnosti slovenščine. Za sprotno sledenje vsem jezikoslovnim področjem bi si tovrstnih priročno oblikovanih zbornikov z obzirnim obsegom in tehtno izbrano vsebino želeli čimveč.

Andreja Žele, Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Novi trg 2, 1000 Ljubljana
E-pošta: andrejaz@zrc-sazu.si