Andreja Trenc

# Splošni korpusi sodobne španščine. Kratek pregled

Korpusno jezikoslovje je disciplina ali metodologija, katere namen je opis, pojasnilo ali razčlenitev jezikovnih pojavov s pomočjo empiričnega jezikovnega gradiva s primeri jezikovne rabe (Zulaica Hernández, 2014, 216). To je razmeroma mlado področje jezikovne analize, ki je nastalo v okviru ameriškega strukturalizma in etnolingvistike na prehodu v drugo polovico dvajsetega stoletja, pravi razmah pa je doživelo zlasti z razvojem tehnološko podprtih obsežnih jezikovnih baz na prelomu tisočletja. Namen pričujoče recenzije je pregled nekaterih vidnejših korpusov španskega jezika kot gradiva za sodobno jeziko(slo)vno analizo, saj se računalniško podprta orodja za znanstveno raziskovanje že uspešno kosajo s tradicionalnimi bibliografskimi viri oziroma zbiri besedil. Prodor korpusov in korpusne analize v jezikoslovje v primerjavi s tradicionalnimi kvalitativnimi opisnimi sredstvi je mogoče ohlapno zarisati v štirih smereh: 1. Računalniška obdelava in mehansko vzorčenje jezikovne rabe s pomočjo velikega nabora razpoložljivih virov v slovaropisju omogoča podrobnejši, izčrpnejši ter interdisciplinarni opis opazovanega jezikovnega pojava v sinhronem preseku, ki nadomešča diahroni nabor pisnih besedil, ter s sistemi jezikovnega označevanja zmanjšuje možnost napake, anahronizmov ipd. zaradi človeškega vpliva pri snovanju in rabi slovarjev ter drugih leksikografskih virov. 2. Zgledi dejanske rabe jezika, kot se je oblikovala v konkretni jezikovni skupnosti s sredstvi opisne slovnice, dopolnjujejo informacije, ki jih je mogoče izluščiti iz normativne slovnice ali slovarja. 3. S kvantitativno analizo empiričnih podatkov, zbranih v večjih splošnih jezikovnih bazah oziroma korpusih, ustrezno premoščajo subjektivne kriterije nabora ali razčlembe, ki jih upoštevajo drugi besedilni zbiri, oblikovani po besedilno-zvrstnem, zgodovinskem ali povsem arbitrarnem kriteriju (antologije ter zbirke knjižnega jezika, oblike spletnega sporazumevanja, poizvedbe s pomočjo spletnih brskalnikov). Korpusni računalniški vmesniki namreč uporabljajo t. i. konkordančnike, ki jezikovno-znanstveno delo z besedili podprejo z ustreznimi metapodatki in označitvami oziroma diatopičnimi, stikovno in družbenozvrstnimi, časovnimi, slogovnimi ter slovničnimi kvalifikatorji. 4. Jezikovne baze naj bi temeljile, kot trdi Biber (1993, 256), na zastopanosti oziroma uravnoteženosti vključenega besedilnega gradiva, na ta način pa naj bi lažje premostile izzive *ad hoc* zbiranja raznovrstnih zapisanih gradiv v stvarnem času ter zagotavljale statistično zanesljivost večjega obsega podatkov.

Zdi se, da v španskem prostoru[1] število raziskav, ki se opirajo na opisane taksonomske možnosti rabe jezikovnih korpusov, podpira relevantnost umestitve v

1 Podroben izbor predstavi Rojo (2016).

sodobno jezikoslovno misel ter njeno uporabo na številnih izbranih področjih zlasti uporabnega jezikoslovja: v besediloslovju, slovaropisju ter analizi besedja, pragmatični analizi in v teoriji diskurza, družbenem jezikoslovju in analizi medijskega diskurza, prevodoslovju ter pri didaktiki (tujih) jezikov.

Ob pregledu splošnih korpusov španskega jezika, s poudarkom zlasti na novejših bazah, ki jih ponujajo vidnejše španske in latinskoameriške jezikoslovne institucije ob podpori svetovnega spleta, kot so CREA, CORDE ter najnovejši CORPES XXI, ne kaže prezreti pomembnih teoretskih predhodnikov, ki so tlakovali pot sodobni analizi korpusnega gradiva. Metodološka izhodišča so se izoblikovala že v šestdesetih letih dvajsetega stoletja v okviru tvorbeno-pretvorbene slovnice (Chomsky, 1965), ki je težišče jezikovne analize z ravni manjših skladenjskih prvin prestavila na raven *makrostruktur*, z opredelitvijo t. i. idealnega naravnega govorca pa je poleg nujne vpeljave jezikovnega opisa na skladenjski in besedilni ravni v središče procesa umestila stvarno rabo jezika v »naravnih« okoliščinah sporazumevanja. Onkraj izhodišč tvorbeno-pretvorbene slovnice, brez katerih bi bila razvojna stopnja številčno in računalniško podprtih sodobnih baz nepredstavljiva, pa je razvoj na prehodu tisočletja trčil ob neogibno teoretsko čer, ki jo zmogljivost tedanjih rudimentarnih tehnoloških vmesnikov ni zmogla prav(očasno) nasloviti. Opis rabe jezikovnega pojava, ki je slonel na intuitivnem zaznavanju, se je upiral znanstvenim metodam, ki temeljijo na zastopanosti ter zlasti preverljivosti pojava z vidika množičnega uporabnika ali celotne govorne skupnosti v konkretnem časovno-prostorskem preseku. Znanstveno uporabnost pri slovnični in besedni analizi so z vključitvijo številčnega ter pestrega jezikovnega gradiva poskušali zagotoviti že predhodniki sodobnih jezikovnih baz.

Med prvimi poizkusi sistematičnega pristopa k analizi zbira besedil v španščini velja omeniti *Diccionario de construcción y régimen de la lengua castellana* kolumbijskega jezikoslovca Rufina Joséja Cuerva (1872, [2]1994), ki je ponudil opis jezikovnega stanja, skladnje in oblikoslovja v ameriški različici španščine na prehodu iz 19. v 20. stoletje na osnovi obsežnega zbira književnih besedil. Kasnejši razvoj korpusov španščine je ponudil interdisciplinarni okvir, v katerega so raziskovalci in uredniki vključili zvrstno pestrejša daljša, a metodološko okrnjena besedila, ki so bila predvsem strokovne oziroma ciljane narave. Med njimi velja omeniti korpus ameriške in evropske španščine Lovaina s 110.000 vzorci 39 besedil različnih avtorjev in področij, napisanih med letoma 1922 in 1988, korpus CUMBRE z dodano frekvenčno porazdelitvijo gesel in besednih zvez, ki je izšel v tiskani obliki v istoimenskem slovarju (Sánchez, 2006), ter nekaj istočasnih projektov, ki so pri opisu stanja v *španskem jeziku* vselej sledili razvoju medmrežja: SOL (*Spanish on Line,* 1998), CRATER (*Corpus Resources and Terminology Extraction,* 1994) z zbirom strokovnega tehniškega izrazoslovja ter večjezični elektronski korpus *Corpus of Contemporary Spanish* (2012), ki je

vključeval pet milijonov vzorcev. Sočasno so v okviru diahronega raziskovanja nabor razpoložljivih korpusov španščine dopolnili tudi projekti elektronskega označevanja zgodovinskih besedil, med katerimi po številu izstopa korpus srednjeveških besedil v španščini *ADMYTE* z javno objavljeno bazo 12 milijonov podatkov iz leta 1991. Vpliv novih tehnologij in zahodnih praks v jezikoslovnem raziskovanju na izoblikovanje podrobnejših kriterijev iskalnih orodij, ki bi vključevala tudi poizvedbe z abstraktnimi jezikovnimi kriteriji, ne le z besedjem in besednimi zvezami, je spodbudil snovanje splošnih korpusov (sodobne) španščine, v prvi različici CREA in CORDE (2013) z nekaj sto milijoni vključenih podatkov obeh zemljepisnih različic španščine od sedemdesetih let prejšnjega stoletja do leta 2000 ter CORPES XXI, ki je leta 2015 nastal na pobudo Španske kraljeve akademije v javno dostopni različici, s kasnejšimi posodobitvami v obsegu primerov ter tehničnimi možnostmi uporabe. Slednje so omogočile zlasti podrobno ter ciljano jezikovno raziskovanje s pomočjo sočasnega upoštevanja več abstraktnih kriterijev, CORPES XXI, ki zajema ustno ter pisno gradivo predhodnih dveh korpusov, pa tudi poizvedbe z izključno slovničnimi kriteriji (jezikovnimi kategorijami in obrazili). Projekt se kosa z drugimi sodobnimi korpusi, ki so relevantni za projektne raziskave posameznih področij španskega jezika: CE (*Corpus del español*), *Proyecto de norma culta* (Samper, 1995) z dialektalnimi različicami dvanajstih zemljepisnih področij ameriške španščine, PRESEEA (*Proyecto de estudio sociolingüístico de España y de* América, 2014), COVJA (*Corpus oral para el estudio del lenguaje juvenil,* 1997), *Corpus de conversaciones coloquiales* raziskovalne skupine Val.Es.Co (2013) Univerze v Alicanteju ter C-Or-DiAL (*Corpus oral didáctico anotado lingüísticamente,* 2012).

Projekt *Corpes XXI* sta Španska kraljeva akademija in Združenje akademij španskega jezika leta 2016 uporabnikom ponudila v javno dostopni obliki. Kot je navedeno na spletni strani avtorjev obravnavane jezikovne baze, je baza v prvi izpopolnjeni različici leta 2016 obsegala 277 milijonov gesel, pridobljenih v letih 2001–2015, v različnih prenosniških zvrsteh (pisni in govorjeni diskurz), znotraj teh zvrsti pa še nekaj besedilnih oblik (umetnostni jezik proze, novinarski jezik, pravni jezik v pisnem diskurzu ter dvogovori, medijski prenosi, razgovori in vsakodnevni pogovori v govorjenem diskurzu). CORPES XXI omogoča iskanje z več funkcijskimi besedilnimi kriteriji hkrati, tako prenosniškimi in socialnimi zvrstmi kot prostorsko narečnimi; vključeni vzorci zajemajo 70 % ameriške španščine ter 30 % evropske španščine, podrobneje pa se delijo tudi na govorna področja oziroma posamezne različice (mehiško, karibsko španščino, španščino območja Río de La Plata ter evropsko španščino). Prva različica jezikovne baze (2015) in kasnejša posodobitev iz leta 2018, ki je nadgradila le obseg gesel v segmentu besedilnih tipologij, v raziskovanje v okviru korpusnega jezikoslovja v španščini vnašata dragocen prispevek ne le v

tehnično-statističnem smislu zaradi večje zanesljivosti in zastopanosti jezikovnih podatkov, temveč in predvsem, kot zagotavljajo avtorji korpusa, v smislu razvoja in širitve možnosti analize podatkov na več področij jezikoslovne analize – od besedja do družbenega jezikoslovja, oblikoslovja in skladnje, iskanje s slovničnimi kriteriji ter poizvedbe z najmanj dvema sočasnima kriterijema v posameznem iskalnem nizu (po oddaljenosti ali zaporedju).

Vpogled v razvoj korpusnega jezikoslovja na področju španskega jezika ponuja bogat nabor sredstev za sodobnega jezikoslovca hispanista, hkrati pa odpira prostor izzivom, ki jim bo ob sedanjih smernicah ter tehnološkem ritmu bržkone kos posodobljena korpusna praksa. Med njimi je zanesljivo vprašanje, kako najti idealno ravnovesje med kvalitativno in kvantitativno analizo obsežnega zbira besedilnega gradiva, ki terja razširitev kompleksnih kombinacij konkordančnikov ter izbirnih iskal za ustrezno pragmatično, semantično ali oblikoslovno-skladenjsko analizo (so) besedila, ki trenutno še ne omogoča poustvaritve nejezikovnih, okoliščinskih vplivov na jezikovne pojave, kot so neverbalna govorica, družbenokritična misel, vpliv govornih dejanj ter implicitnega konteksta, metaforičnega jezika itd. Po drugi strani pa bi prav raziskovanje analognih, abstraktnih pomenskih jezikovnih pojavov odprlo vprašanje, kako zagotoviti ustrezno statistično zanesljivost ter objektivnost opisa, ki ga doslej omogoča digitalna obdelava. Ob odgovorih na zapisane izzive, ki jih bo podal prihodnji razvoj, prispevek razvoja korpusov sodobne španščine v dostopni obliki poleg taksonomije in znanstvene presojnosti nedvomno pomeni razsrediščenje tradicionalnih opisnih metod z usmeritvijo k uporabni analizi, ki upošteva tudi množičnega uporabnika. To pa sta podporna stebra sodobne informacijske družbe.

## Bibliografija

### Viri

Izbor korpusov sodobne španščine:

CE: Corpus del español [www.corpusdelespanol.org/].

CORDE: Corpus diacrónico del español [http://rae.es/recursos/banco-de-datos/corde].

C-Or-DiAL: Corpus oral didáctico anotado lingüísticamente [http://lablita.dit.unifi.it/corpora/cordial].

CORPES XXI: Corpus del español del siglo XXI [http://rae.es/recursos/banco-de-datos/corpes-xxi].

Corpus of Contemporary Spanish [http://spanishfn.org/tools/cea/english].

COVJA: Corpus oral de la variedad juvenil universitaria del español de Alicante; integrado en el Corpus oral para el estudio del lenguaje juvenil y del español hablado en Alicante.

CRATER: Corpus Resources and Terminology Extraction [http://ucrel.lancs.ac.uk/projects.html#crater].

CREA: Corpus de referencia del español actual [http://rae.es/recursos/banco-de-datos/crea].

PRESEEA: Proyecto para el estudio sociolingüístico del español de España y de América [preseea.linguas.net/].

Literatura

Chomsky, N., *Aspects of the Theory of Syntax,* Massachusetts 1965/1969, MIT Press.

Biber, D., Representativness in corpus design, *Literary and linguistic computing* 8/4, 1993, str. 243–257.

Rojo, G., Corpus textuales del español, v: *Enciclopedia de lingüística hispánica* (ur. Gutiérrez Rexach, J.), London 2016, str. 285–295.

Cuervo, R. J., *Diccionario de construcción y régimen de la lengua castellana por Rufino José Cuervo*, continuado y editado por el Instituto Caro y Cuervo, Santafé de Bogotá 1872 $^{2}$1994.

Sánchez, A., *Diccionario CUMBRE,* Madrid 2006.

Zulaica Hernández, I., Lingüística de corpus, v: *Enciclopedia de lingüística hispánica* (ur. Gutiérrez Rexach, J.), London 2016, str. 216–224.