FILOLOGIJA 30–31, Zagreb 1998

UDK 808.62-31
Izvorni znanstveni članak
Primljen 12.V.1998.
Prihvaćen za tisak 15.VI.1998.

Maja Bratanić
Fakultet prometnih znanosti
Vukelićeva 4, HR-10000 Zagreb

KORPUSNA LINGVISTIKA NA KRAJU 20. STOLJEĆA I IMPLIKACIJE ZA SUVREMENU HRVATSKU LEKSIKOGRAFIJU

Korpusna lingvistika i na njoj utemeljena empirijska istraživanja jezika bilježe posljednjih desetljeća uspon i domete po kojima umnogome nadmašuje druge suvremene lingvističke discipline. Osobito su se dalekosežnima metode korpusne lingvistike pokazale u području leksikografije. Primjena nove informacijske tehnologije u lingvistici se potpuno oslanja na jezične korpuse. Autorica upozorava na neodgodivu potrebu stvaranja hrvatskoga nacionalnoga korpusa i raspravlja o nekim pitanjima bitnima za njegovo valjano planiranje i djelotvorno korištenje.

Jedno je od najakutnijih otvorenih pitanja hrvatske leksikografije — usuđujem se reći alarmantan zahtjev što se postavlja suvremenomu hrvatskomu jezikoslovlju — stvaranje korpusa suvremenoga hrvatskoga jezika koji bi mogao odgovoriti potrebama modernoga lingvističkoga istraživanja, a koji bi, dakako, bio u skladu s današnjim tehnološkim mogućnostima informatičke industrije granice kojih se pomiču praktički danomice. Dopuštam sebi pomalo pristran stav, govoreći o toj temi, jer sam imala zadovoljstvo sudjelovati — sada se to posebno dobro uočava — u tada pionirskom projektu izgrađivanja prvoga u lingvističke svrhe sustavno koncipiranoga računalnoga korpusa hrvatskoga jezika što su ga početkom sedamdesetih godina pokrenuli Milan Moguš i Željko Bujas u Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu.[1] Ta je inicijativa tek nekoliko godina zaostajala za kapitalnim događajem u rađanju suvremene korpusne lingvistike (tzv. Brownovim korpusom suvremenoga američkoga engleskoga jezika) koji je označio početak jedne lingvističke struje što je iz pozicije gotovo subverzivne postbloomfieldijanske djelatnosti u odnosu na dominantnu chomskijansku lingvistiku do

[1] Ovom prigodom nećemo podsjećati na neke druge oblike korpusnih istraživanja u nas, isto tako avangardna, a koja se vezuju uz djelovanje akad. Rudolfa Filipovića i suradnikā na kontrastivnom istraživanju engleskoga i hrvatskoga jezika.

današnjih dana prerasla u samosvojnu maticu i za sobom povukla cijeli spektar najrazličitijih istraživanja kojima bi se kao zajednički nazivnik mogao označiti oslonac u empirijskom istraživanju, u konkretnoj, živoj jezičnoj građi koja se često ne ponaša u skladu s očekivanjima lingvista teoretičara (usp. Teubert 1995, 1996b).

Tako su se razvila i sustavno koegzistiraju dva komplementarna pogleda na jezik (a danas se čini da je gotovo i moralo biti tako jer je introvertirana lingvistika zasnovana na introspekciji morala postati pretijesnom): onaj faktografski koji iz stvarnoga jezičnoga fenomena vodi generalizacijama, te onaj koji polazi od univerzalija i potkrepljuje ih lingvističkim konstruktima. Drugim riječima, pristup koji polazi od jezika kao mentalne činjenice suprotstavljen je onomu koji ga prije svega vidi kao društvenu datost i zanima se za njegovu komunikacijsku funkciju. Metaforički rečeno, u izvrnutoj bi se i ponešto isforsiranoj perspektivi, moglo reći da korpusi u određenom smislu isto tako simuliraju lingvističku kompetenciju idealnoga izvornoga govornika u rasponu koji može nadići pretpostavke bilo kojega konkretnoga modela. Naime, suvremeni korpusi sastavljeni od golemoga broja potvrđenih tekstova (danas već s više stotina milijuna riječi) omogućuju empirijska istraživanja donedavno doslovno nezamislivih razmjera. Posljednje dvije dekade ovoga tisućljeća obilježene su punim cvatom korpusne lingvistike, i njezinim prerastanjem iz metodologije u nov pristup promatranju jezične građe iz kojega proizlazi i nova vrsta znanja o jeziku (usp. Leech 1992).

U devedesetima već dolazi i do svojevrsnoga raslojavanja i specijalizacije korpusa jer se oni počinju izrađivati u strože ciljane svrhe, dok tzv. *standardni korpusi*, tj. korpusi općega jezika, postaju općim mjestom i nezaobilaznom pomoćnom aparaturom svake razvijene jezikoslovne struke.

Svrhovitost, pa i nužnost porabe korpusa u leksikografskom radu, dakle na obradbi leksika, kao ishodišne ili pomoćne aparature, danas nije više potrebno braniti. Dosadašnja su istraživanja u području korpusne lingvistike višestruko raspršila svaku sumnju u opravdanost takva pristupa (ne samo na razini svjetski poznatih projekata već i u okvirima naših ograničenih istraživanja, usp. npr. Bratanić 1993). Teško je ne složiti se s maksimom što je jedan od pionira korpusne leksikografije, John Sinclair, često ističe i parafrazira: da odgovor na sva pitanja o jeziku jest u jeziku samom te da prema tomu najbolje potkrepe svakom leksikografskom objašnjenju leže u stvarnim primjerima.

Ne treba zanemariti ni to da su se različiti oblici "korpusa" rabili u dugostoljetnoj tradiciji hrvatske leksikografije, jer bez nekog oblika rječnika građe teško da može biti i leksikografije, no dobro je upozoriti i na neke bitne razlike. Na tradicionalan način prikupljena i na karticama ispisana zbirka potkrepa i danas čini solidnu osnovu svakom ozbiljnom leksikografskom projektu (osobito jednojezičnom), no metode i instrumentarij suvremene kor-

pusne lingvistike taj su pristup razvile do nepredvidljivih granica. Dokazalo se, naime, da dok tradicionalna metoda u načelu omogućuje solidan uvid u širinu jezika najčešće uspješno zahvaćajući raspon manje čestih riječi ili značenja, suvremeni korpusni pristup nema alternative u omogućivanju dubinskoga uvida u jezik, jer se samo promatranjem velikoga broja pojava iste leksičke jedinice može pouzdano saznavati o njezinu gramatičkom ponašanju, kolokacijskom potencijalu i drugim osobitostima sintagmatske prirode.

Polazeći od činjenice da smo danas, u odnosu na spomenutu situaciju od prije dvadeset i pet godina, u znatnom zaostatku ne samo za velikim svjetskim jezicima, skrenut ću pozornost na neke bjelodane činjenice koje valja imati na umu pri stvaranju nacionalnoga korpusa hrvatskoga jezika, tom prioritetnom jezikoslovnom zadatku kojem je danas u nas okrenuto više institucija.

Najočitija je tendencija u razvoju korpusa rast njihove veličine. Korpusi prve generacije — npr. jednomilijunski engleski korpusi Brown ili LOB — danas se nazivaju korpusima uzorcima. Drugu generaciju, u rasponu između sedam i tridesetak milijuna pojavnica, činili su korpusi poput Sinclairova *Birmingham Collection of English Texts*, dok su korpusi treće generacije sastavljeni od stotina milijuna riječi i najčešće nastaju kao nusproizvodi suvremenih elektroničkih komunikacijskih sustava.

U veličini se kriju stanovite moguće zamke (Leech 1991:11–12) jer sama veličina korpusa ne jamči i srazmjernu raznovrsnost građe, a s druge strane, vrijednost i "dubina" rezultata istraživanja korpusa izravno ovise o razvijenosti programskih alata za njegovo pretraživanje i lingvističku raščlambu.

Zbirka tekstova sama po sebi nije korpus. Korpus mora biti, kako se to obično kaže, **reprezentativan**. Najopćenitije rečeno, korpus se može smatrati reprezentativnim ako se nalazi što ih on omogućuje mogu smatrati općenito valjanima za neki zamišljeni veći korpus ili jezik u cjelini. S druge strane korpus uvijek predstavlja samo tekstove zastupljene u njemu, a ne lingvistički univerzum (Teubert 1995:119). Kriteriji će, ističe Teubert, biti impresionistički sve dok se ne budu mogli osloniti na statističke ili neke druge modele. Danas su zahtjevi u pogledu reprezentativnosti korpusa razumniji i realniji pa se, kad je riječ o leksiku, realnom smatra težnja za stvaranje tzv. zasićenoga korpusa. Korpus je zasićen onda kad stopa rasta novih riječi u odnosu na povećavanje korpusa postane stalna (Teubert 1995:119–120).

John je Sinclair razradio tipologiju korpusa koja može biti od pomoći pri koncipiranju korpusa (Sinclair 1995; Teubert 1995):

Specijalni korpusi — korpusi su s posebnom namjenom. Upravo stoga nisu uravnoteženi i ne mogu služiti u svrhu drugu osim one kojoj su namijenjeni jer u protivnom mogu dati iskrivljenu sliku. Prednost im je u odnosu na uravnotežene korpuse u tome što se zahvaljujući ciljanomu izboru građe fenomen koji se istražuje ovdje javlja učestalije čak i onda kad su znat-

no manji od tzv. uravnoteženih korpusa.

Referentni korpusi najbliži su sada već zastarjelu pojmu reprezentativnoga korpusa, a temelje se na nekim relevantnim parametrima oko kojih je postignut lingvistički dogovor te obuhvaćaju i pisani i govoreni jezik, formalne i neformalne njegove registre itd. Trebali bi poslužiti najraznolikijim namjenama kompenzirajući potrebu za specijalnim korpusima. Prema njima se postavljaju standardi za leksikone te služe kao poligoni za razne programske alate i različite posebne primjene. Minimalna im je veličina oko pedeset milijuna riječi, a europski se standardi danas ustaljuju na oko sto milijuna riječi.

Kontrolni korpus (pojam što ga je i u operativnom smislu uveo sam Sinclair u radu na projektu Collins–COBUILD) označava korpus kojem je svrha neprestano ažuriranje referentnoga korpusa kako bi on mogao odražavati jezične mijene, čuvajući pri tom njegovu ravnotežu i sastav. Takav se korpus izrađuje obično za godinu dana, a u skladu s modelom referentnoga korpusa.

Oportunistički korpusi manje su skupe alternative referentnim korpusima. Danas se popularno nazivaju arhivima i zapravo su zbirke svih tekstova u elektroničkom obliku do kojih se može doći na najjeftiniji način. Oni su u određenom smislu virtualni korpusi, a stvarni se korpusi mogu izlučiti iz njih prema zahtjevima konkretnoga projekta.

Za višejezična su istraživanja pa tako i za višejezičnu leksikografiju danas potrebne dvije vrste korpusa:

Usporedivi korpusi jesu korpusi različitih jezika sastavljeni prema identičnom obrascu. Postaju nezaobilaznima pri izradi dvojezičnih i višejezičnih leksikona te novu generaciju rječnika.

Usporedne (paralelne) korpuse čine tekstovi na jednom jeziku i njihovi prijevodi na drugi jezik, a namijenjeni su ponajprije uspostavljanju prijevodne ekvivalencije. U tu svrhu moraju biti posebno grafički priređeni (Teubert 1995:121, 1996a; Bratanić 1997).

Te su se dvije vrste korpusa osobito stimulativnima pokazale u novom pristupu strojnom prevođenju i izradi tomu namijenjenih alata (Sinclair 1996a, 1996b).

Poseban je sklop problema vezan uz sastavljanje korpusa i to čini zasebnu temu. Sinclair (1991:13) iznosi zanimljivu misao da bi se određivanjem sadržaja korpusa, u smislu izbora tipova i proporcija građe, trebali baviti primarno sociolozi kulture, dok je zadatak lingvista da opisuje i analizira sam uzorak jezika. U pionirskim fazama korpusne leksikografije tim će se poslom nužno baviti lingvisti, no dobro je razmotriti mogućnost uključivanja i drugih društvenih znanosti.

Korpusi su još uvijek uglavnom reprezentativni samo za medij pisanoga jezika pa se, kad govorimo o toj temi, kao i u ovom slučaju, oni najčešće i podrazumijevaju. Put do korpusa govorenoga jezika vrlo je mučan, počevši

od načina prikupljanja građe do njezine transkripcije u pisani oblik koji omogućuje da se sačuvaju svi podaci relevantni za analizu. U ovom je slučaju, dakako, i mnogo teže odlučiti što će se smatrati "reprezentativnim", no hrvatsko jezikoslovlje morat će se u doglednoj budućnosti poduhvatiti i projekata koji bi vodili tomu cilju.

Sljedeće pitanje što ga valja dotaknuti odnosi se na standardizaciju korpusa, tj. na način njihova ujednačena označavanja (ili kodiranja) u svrhu čuvanja izvornih podataka o tekstovima koji ga sačinjavaju. Standardiziranje korpusa u današnje vrijeme osobito dobiva na važnosti i zbog interjezičnih istraživanja za koja je podudaranje različitih tekstovnih kategorija vrlo relevantno zbog mogućnosti njihova uspoređivanja i uspostavljanja međusobnih veza (Teubert 1995:122). Odluke u vezi s označavanjem korpusa na prvoj se razini odnose na različite grafičke elemente teksta — konkretno: ime(na) autora, naslove, podnaslove, odlomke, potpise pod slikama i drugim grafičkim prilozima, bilješke, navode, tip slova itd. — i one se, kad se jednom ustanove, moraju vrlo jednoznačno i dosljedno provoditi u korpusu kao cjelini jer će o tome izravno ovisiti i stupanj njegove iskoristivosti na tom planu. Jedino se na taj način može, primjerice, sa sigurnošću ustvrditi da se neka sintagma ili frazem javlja poglavito u naslovima novinskih tekstova, dok za drugi tip diskurza uopće nije tipična. S porastom veličine korpusa takvi podaci postaju sve važnijima, a na sastavljače korpusa postavljaju dodatne zahtjeve. Stoga je za potrebe univerzalnog jednoznačnog označavanja tekstova tijekom osamdesetih razvijen i široko prihvaćen takozvani standardni uopćeni jezik za označavanje (SGML), a u novije se vrijeme naporima za ujednačeno označavanje korpusa različitih jezika pridružila i tzv. Inicijativa za kodiranje tekstova (TEI) s osnovnim ciljem da oni budu što sumjerljiviji i lakše razmjenjivi u elektronskoj formi.

Pri planiranju i izradi korpusa može se, nadalje, odlučiti da se korpusu ili pojedinim njegovim dijelovima, pridruže i drugi, striktno lingvistički podaci na morfološkoj ili/i sintaktičkoj razini. Takve će odluke primarno ovisiti o izravnoj namjeni korpusa, no Teubert ispravno ističe (1995:124) da se apriornim nametanjem tradicionalnih kategorija elementima korpusa mogu ograničiti ili čak onemogućiti novi lingvistički uvidi.

Kodiranje korpusa može se, opet u skladu s ciljevima namjeravanoga istraživanja, zamisliti i provesti i na semantičkoj razini te osobito na razini diskursa.

Nov niz problema koje valja anticipirati pri izradi i daljnjoj obradbi korpusa u leksikografske svrhe javit će se pri lematizaciji korpusa. To možemo potkrijepiti i osobnim iskustvom u radu na dosad jedinomu tako obrađenu korpusu hrvatskoga jezika, izrađenu u Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu — gdje je pri postupku lematizacije došla do izražaja potreba razrješavanja problema i na razinama iznad puke morfologije, što se u načelu

moglo i predvidjeti, ali ne u pravoj mjeri (Moguš et al., u tisku). Lematiziranje korpusa preduvjet je za učinkovitiju leksikografsku analizu, no za velike korpuse od više desetaka ili čak stotina milijuna riječi taj postupak može biti prezahtjevan. Rješenje će valjati potražiti u isprva poluautomatskoj, a potom i u većim dijelom automatskoj lematizaciji korpusa što će je omogućiti postojeća i buduća istraživanja u području automatske morfološke analize i generiranju hrvatskoga jezika. Uza sve to, i dalje će sastavljačima često preostati konačna odluka o tome koji oblik valja izabrati za lemu, moguće nedoumice oko distribucije značenja između različitih oblika leme itd. (usp. Sinclair 1991:41).

Usporedo s razvojem i izgradnjom nacionalnoga korpusa hrvatskoga jezika, moraju se razvijati i odgovarajuća programska podrška i alati za analizu korpusa. Bogatstvo podataka što ga nudi korpus može djelovati kaotično. Da bi se to jezično blago moglo dosljedno otkrivati i izdašno crpsti, potrebno mu je prići sustavno, s razrađenim instrumentarijem. Strojna se oprema razvija znatno brže nego programska, a da bi se adekvatna oruđa za lingvističku analizu mogla razraditi tako da iznesu na vidjelo relevantne osobitosti korpusa, nužno je analizirati sam korpus. Riječ je o svojevrsnom paradoksu rada s korpusom. Analiza korpusa podrazumijeva jasno formuliranje leksikografskih kategorija. Ona je stoga plodotvorna u oba smjera. Korpusna je lingvistika razvila metode pomoću kojih se mogu identificirati i opisati leksičke jedinice od razine riječi do složenih leksičkih sklopova, a hrvatska lingvistika danas mora naći način kako da ta znanja primijeni u istraživanju i opisu vlastitoga jezika. Alati za analizu korpusa još su razmjerno primitivni, što u krajnjoj konsekvenciji i pri analizi korpusa ostavlja dosta prostora intuiciji. To nas vraća ishodišnim razmišljanjima o poziciji dviju dominantnih, prividno sasvim suprotstavljenih lingvističkih škola. Stavovi su se u međuvremenu s obje strane umnogome međusobno približili.

Literatura

Bratanić, Maja. 1993. Leksikološko-leksikografski potencijal jednomilijunskoga korpusa hrvatskoga književnog jezika. *Bilten Zavoda za lingvistiku* 6, 17–21.

Bratanić, Maja. U tisku. Od intuicije do opservacije i nazad (višejezična leksikografija i paralelni korpusi). *Suvremena lingvistika*.

Leech, Geoffrey. 1991. The state of the art in corpus linguistics. *English Corpus Linguistics*, K. Aijmer i B. Altenberg (eds.). London and New York : Longman. 8–29.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. U J. Svartvik (ed.), *Directions in Corpus Linguistics*, Proceedings of Nobel Symposium 82, Stockholm 4–8 August 1991. Berlin : Mouton, de Gruyter.

Moguš, Milan, et al. U tisku. *Čestotni rječnik hrvatskoga jezika*. Zagreb : Školska knjiga.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Sinclair, John, and J. Ball. 1995. Text Typology (External Criteria). Draft version. Dokument u elektroničkom obliku, Pisa EAGLES ftp server, Birmingham.

Sinclair, J. M. et al. 1996a. Corpus to Corpus: A Study of Translation Equivalence. Predgovor u *International Journal of Lexicography* 9:3, 171–178.

Sinclair, J. M.1996b. Multilingual Databases. An International Project in Multilingual Lexicography. *International Journal of Lexicography* 9:3, 179–196.

Svartvik, Jan. 1992. Lexis in English language corpora. *Euralex '92 Proceedings*, Part I, Tampere : University of Tampere. 17–31.

Teubert, Wolfgang. 1995. Language Resources: The Foundations of a Pan-European Information Society, *Proceedings* of the First Trans-European Language Resources Infrastructure (TELRI) Seminar *Language Resources for Language Technology*. 105–128.

Teubert, Wolfgang. 1996a. Comparable or Parallel Corpora?. *International Journal of Lexicography* 9:3, 238–264.

Teubert, Wolfgang. 1996b. Editorial. *International Journal of Corpus Linguistics* 1:1, III–X.

Corpus linguistics at the end of the 20th century and the implications for contemporary Croatian lexicography

Summary

Corpus based linguistics has in the decades following the early sixties gradually extended its scope and influence and has practically become a dominant linguistic mainstream itself. Particularly far reaching has been its impact in the field of lexicography. New language technologies are entirely based on large language corpora. The paper pleads for the urgent development of a national Croatian corpus and discusses a number of issues relevant to corpus planning and exploitation.

Ključne riječi: korpusna lingvistika, hrvatska leksikografija
Key words: corpus linguistics, Croatian lexicography