KORPUSNI PRISTUP SOCIOLINGVISTIČKOM ISTRAŽIVANJU

Nikola Dobrić*

Sažetak: *Jezik se nalazi u samom centru društva, društvenog delovanja i poimanja sveta i kao takav treba da predstavlja početnu stepenicu u svakom humanističkom istraživanju. Sociolingvistika poprima novu dimenziju kada se spoji sa prednostima koje pruža korpusna metoda jezičke analize, a samim tim i društva. Ovaj rad će predstaviti pojam korpusa i korpusne analize jezika kao centralne osnove svake teorijske postavke o društvu i poimanju sveta oko njega.*

Ključne reči*: korpus, frekvencija, sociolingvistika, sociolingvitička promenljiva, statistika*

Abstract: *Language is at the very centre of society and the way it comprehends the world around it, and as such should be seen as the starting point for any humanities-based research. Sociolinguistics takes on a new dimension when combined with the advantages provided by the corpus method of linguistic analysis. This paper aims to present the notion of corpora as the centre of every theoretical research into society and the way it sees the world.*

Key words: *corpus, frequency, sociolinguistics, sociolinguistic variables, statistics*

Bliska veza jezika i društva

Značenje jedne reči je parafraza, a svaki je jezik stoga samoreferentni sistem. Prirodni jezik u svakodnevnoj upotrebi se naziva *diskurs* i predstavlja sistem u kome je funkcionisanje unutrašnjih elemenata određeno ne zakonitostima spoljnog sveta nego dogovorom i pregovorom oko značenja unutar samog jezika. Kada razgovaramo, jezik se ne poziva u svom značenju na spoljašnju stvarnost već na ono što je prethodno rečeno unutar jednog *diskursnog društva*, odnosno unutar društva koje koristi određeni jezik. Zbog toga se značenje u jeziku uvek pregovara i uvek oslanja na proteklo društveno i civilizacijsko iskustvo. Svaka leksička jedinica zavisi od neke druge leksičke jedinice i odnosi se na neku drugu leksičku jedinicu, a sve zajedno na kraju na celokupnu ljudsku komunikaciju i jezičku i kulturološku tradiciju. Jezik se sastoji samo od svedočenja jednog društva koje ga koristi o svetu koji okružuje to društvo. Jezik nije iskustvo realnog sveta iz prve ruke. Jezik ne predstavlja realnost već se njegov odnos sa realnim referentnim objektima ostvaruje pomoću veze koju uspostavlja zajednica koja koristi dati jezik sa stvarnim svetom i pojavama unutar njega. Da je jezik stvarni prikazatelj realnosti, ne bi bilo slučajeva da na primer u nemačkom jeziku reč za boju *narandžasto* nije postojala sve do 19. veka bez obzira na to što je data boja očigledno teorijski bila prisutna. Ljudi su u tome jedinstveni – jedini su sposobni od svih živih bića da pregovaraju oko značenja i sadržaja svog sveta i da potom dalje određuju značenje na osnovu prethodno dogovorenih pojmova.

Da bi jezik imao značenje potrebno je ispuniti još jedan uslov – jeziku je potrebna publika. Bez publike komunikacija nema smisla. Jezik je pre svega kolektivna pojava. Ne postoji privatni jezik.

Jezik takođe zavisi i od situacije i okruženja. Komunikacija je često deo drugih aktivnosti, kao na primer šetanja u parku ili gledanja televizije. Ona zavisi od konteksta, bilo trenutnog ili društvenog. Značenje se dešifruje pronalaženjem prave parafraze u okviru postojećeg društvenog skupa svedočenja o realnom svetu koje odgovara datom kontekstu.

* Nikola Dobrić, nastavnik engleskog jezika, Visoka poslovna škola strukovnih studija, Novi Sad, Srbija

Ovakva samoreferentnost jezika i oslanjanje na društveno iskustvo pri pregovaranju značenja čini jezik najboljim pokazateljem odnosa jednog društva prema stvarnom svetu i prema samome sebi.

To čini jezik možda najboljom platformom za istraživanje pogleda i ideja koje jedno društvo ima o sebi i o svom okruženju.

Frekventnost kao osnovni parametar

Sociolingvistika je upravo ta naučna osobina koja se bavi zakonitostima odnosa društva i jezika. Osnovni podaci koje jedan jezik pruža u okviru nekog sociolingvističkog istraživanja koji čine bazu sociolingvistike su frekventnosti reči, odnosno koliko se puta neka leksička jedinica ponavlja u okviru određenog konteksta i kako na to utiču razne *sociolinvističke promenljive*. Na primer, možemo brojati koliko se puta određeni gramatički nastavak koristi pri komunikaciji, pa na osnovu takvih podataka pokušati predvideti morfološku upotrebu analiziranih reči u funkciji sintaksičkog okruženja, društvene grupe ili geografskog položaja. Možemo na primer posmatrati jezik sa fonološkog stanovišta i istraživati koliko puta se određena reč izgovara na različite načine, pa na osnovu toga teoretisati o uticaju socioekonomskog statusa, pola, geografske oblasti ili neke druge sociolingvističke promenljive na određeni izgovor reči.

Kakav god pogled na sociološku analizu da izaberemo, istraživanje će se zasnivati na pronalaženju shema i zakonitosti u brojkama frekventnosti date sociolingvističke promenljive.

Definicija korpusa

Da bi bi bilo moguće istraživati ljudsko društvo preko jezika kojeg to društvo koristi potrebno je naći način kako analizirati jezik u svom prirodnom upotrebnom obliku koji bi kao takav predstavljao jedno društvo. Takav skup prirodnog jezika u svojoj stvarnoj upotrebi u celini ili jednog određenog dela tog jezika predstavljen u tekstualnom obliku naziva se korpus.

Korpus u svom osnovnom značenju predstavlja bilo kakav skup teksta, bilo pisanog ili govornog. Knjige koje posedujemo i stoje nam na polici su korpus. Pismeni zadaci i ispiti su korpus. Međutim, korpusi u onom smislu relevantnom za šire sociolingvističke analize predstavljaju nešto složenije konstrukcije. Radi se o opsežnim skupovima jezika, odnosno tekstualnog prikaza jezika, koji se mogu pronaći u digitalnom obliku organizovani u velike računarske baze. Takve korpuse obično sakupljaju univerzitetska tela ili slične institucije. Svaki se takav korpus sastoji od više miliona reči iz različitih jezičkih i društvenih izvora: fikcijskih i dokumentarnih pisanih izvora, akademskih radova, novinskih članaka, telefonskih razgovora, reklama, naučnih predavanja, javnih govora, televizijskih emisija, i slično. Od velike je važnosti za teoretsku vrednost rezultata da tekstovi uvršteni u jedan korpus budu i iz pisanih i iz govornih izvora i da predstavljaju što širu paletu registara, dijalekata i sociolekata. Stoga sledi da jedan takav korpus idealno obuhvata sve moguće pojave jednog jezika upotrebljene u realnim životnim situacijama i kao takve uhvaćene u vremenu i pretočene u digitalni tekstualni oblik.

Postoji veliki broj sociolingvistički vrednih digitalnih korpusa većine svetskih jezika kojima se može pristupiti preko interneta. Što se tiče situacije kod nas, trenutno postoji samo nekoliko većih i značajnijih korpusa srpskog jezika, čiji je kvalitet i upotrebljivosti upitne vrednosti.

Pošto se takvi korpusi dele na *specijalizirane* (koji se sastoje samo od posebno odabranih tekstova, na primer samo od novinskih članaka ili samo od jezika poslovne komunikacije i koji su kao takvi posebno skrojeni za određena istraživanja) i *opšte* (koji treba da predstavljaju jedno diskursno društvo u celini u što širem opsegu njegove komunikacije) bitno je za sve različite vrste sociolingvitičke analize odabrati korpus opšteg karaktera kako bi osigurali sociološku relevantnost i referentnost povratnih informacija.

Reprezentativnost jednog korpusa opšteg karaktera, a samim tim i kvalitet rezultata koje taj korpus pruži prilikom neke analize, vrednuje se ne veličinom nego prvenstveno raznolikošću koja se dobija pravilnim i planiranim odabirom što šireg spektra izvora pri konstrukciji korpusa.

Dakle, korpusi predstavljaju jezik i sva njegova značenja zabeležena u tekstovima pogodnim za analizu. Da bi korpus mogao pružiti pouzdane rezultate, to jest da bi se na osnovu korpusa kao opseg jednog istraživanja moglo postaviti jedno društvo koje ga koristi, potrebno je koristiti korpus dovoljno reprezentativan po pitanju varijacija komunikativnih izvora (registara, sociolekata, dijalekata, pisanog i govornog jezika).

Upotreba korpusa u sociolingvistici

Jedno uspešno istraživanje društva pomoću velikih skupova teksta, odnosno korpusa, koji predstavljaju prirodni jezik u prirodnoj upotrebi sastoji se od nekoliko koraka:

- **Definisanje problema –** sociološka i lingvistička pojava koju želimo da istražimo podrobno se definiše zajedno sa svim relevantnim parametrima i implikacijama;
- **Izbor referentnog korpusa –** cilj je pronaći dovoljno referentan korpus opšteg karaktera koji ima dovoljan broj reči i dovoljan broj varijacija jezika da bi predstavljao jedno društvo u dovoljno reprezentativnom stepenu za ovakvo sociolingvističko istraživanje. Mora se naglasiti da situacija nije jednostavna po pitanju korpusa srskog jezika. Dok postoje desetine onlajn korpusa engleskog jezika, raznih veličina i upotrebljivosti, postoji samo nekoliko takvih korpusa srpskog jezika od kojih nijedan nije zadovoljavajućeg kvaliteta prema parametrima opštosti.[1] Osim digitalnih onlajn korpusa mogu se koristiti i lični korpusi, ali se mora obratiti pažnja na njihovu reprezentativnost;
- **Sirove frekvencije i normiranje –** glavni parametar zakonitosti unutar jednog korpusa, a samim tim i zakonitosti unutar jezika i društva, je frekventnost referentnih izraza uzetih kao sadržaj analize. Frekventnost je definisana kao broj koji predstavlja koliko se puta neki izraz pojavljuje u određenom korpusnom kontekstu što posle dozvoljava izvođenje teorijskih zaključaka vezanih za cilj istraživanja. Podaci dobijeni analizom korpusa nazivaju se *sirove frekvencije* analiziranih izraza. To su neobrađene brojke koje predstavljaju koliko se puta neka reč pojavila u određenom korpusu koji koristimo kao podlogu za istraživanje. Pre nego ti podaci postanu iole pogodni za analizu potrebno je da prođu kroz još nekoliko postupaka. Prvi postupak koji je neophodno primeniti je *normiranje* rezultata. Pošto je u većini slučajeva kada se koriste različiti korpusi broj izraza u njima nejednak, bitno je osigurati da se rezultati frekvetnosti svedu na vrednosti koje bi imale da su korpusi jednake veličine. Proces normiranja ili normalizacije upravo to omogućuje. Korišćenjem jednostavnog matematičog proračuna dobijamo *normalizovane frekvencije* analiziranih izraza;[2]
- **Statistička vrednost rezultata –** sledeći korak u ovakvoj analizi obično nije veoma popularan, a sastoji se od statističke verifikacije dobijenih normalizovanih rezultata. Statističko verifikovanje se sastoji od mnoštva komplikovanijih matematičkih i statističkih metoda [3], za koje na sreću postoji i pomoćni softver, a za cilj ima da potvrdi da li su dobijene vrednosti frekventnosti rezultat nasumične pojave u jeziku ili su stvarno jezički i naučno relevantni. Statistički proračuni se koriste i kod izračunavanja verovatnoće pojave jezičke jedinice u pretpostavljenom diskursu teorijski neograničene veličine. Jednom kada potvrdimo da su podaci statistički vredni, spremni su da se na osnovu njih donesu određeni zaključci vezani za postavljeni cilj istraživanja;
- **Distribucija frekventnosti –** podaci koji prođu kroz sve opisane postupke i uspešno zadovolje tražene statističke parametre matematičke i naučne relevantnosti se zatim predstavljaju u obliku procentualnih vrednosti kroz grafikone. Na osnovu tako predstavljenih podataka moguće je analizirati i razmatrati sociološke i jezičke razloge za dobijenu distribuciju i iz toga izvući određene relevantne zaključke.

[1] *http://www.serbian-corpus.edu.rs/indexns.htm*
http://korpus.matf.bg.ac.yu/prezentacija/korpus.html
[2] N. Dobrić, *Diskurs diskriminacije u engleskom i srpskom*, Filozofski fakultet, Novi Sad, master rad, 2008.
[3] K. Johnson, *Quantitative Methods in Linguistics*. Blackwell Publishing, Oxford, 2008.

Budućnost

Avenije istraživanja koje otvara ovakav pristup očigledno su veoma zanimljive jer svako relevantno teoretisanje fokusirano na mnoge aspekte društva može biti čvrsto poduprto opipljivim dokazima u obliku statističkih podataka koji su društveno reprezentativni i relevantni. Jedini problem i jedina prepreka široj upotrebi ovakvog zanimljivog pristupa istraživanju jezika i društva je nepostojanje dovoljno sociološki relevantnog korpusa srpskog jezika. Svi trenutno postojeći korpusi srpskog jezika pokazuju određene nedostatke u sociolingvističkom smislu. Na primer, korpus srpskog jezika analiziran u opisanom primeru, iako ima zavidan broj reči (dvadeset i četiri miliona), mane prikazuje u pogledu izvora iz kojih je korpus sačinjen koji nisu dovoljno širokog spektra. Dvadeset i dva miliona izraza je dobijeno iz tekstova preuzetih iz dnevnih novina *Politika*, pa je očigledno kako su mnogi oblici upotrebe jezika (kao književni jezik, studentski ispiti, kafanski razgovori, i slično) nepostojeći. Još jedna velika mana je i to što se korpus sastoji samo od pisanog jezika, što je veliki nedostatak jer je govor ipak osnovna prirodna jezička forma. Sve ove mane čine ovaj korpus nepodobnim da se na osnovu njega može pretpostaviti da predstavlja celokupno srpsko društvo.

Zbog toga, i zbog mnogih drugih tehničkih razloga, neophodno je sastavljanje novog, sveobuhvatnijeg korpusa srpskog jezika, na nacionalnom nivou, koji bi obuhvatio i govorni i pisani jezik iz što je moguće više različitih izvora jezika i koji bi konstantno bio održavan i proširivan novim jezičkim pojavama na nivou državnog projekta. Dok se to ne dogodi, postojeći korpusi, koji iako možda ne zadovoljavaju sve potrebne parametre, moraju zadovoljiti kao jedina dostupna podloga za sprovođenje jednog od najvažnijih polja naučnog istraživanja zakonitosti odnosa jezika i društva.

Literatura

[1] Biber, D., Conrad, S., Reppen, R., (2000) *Korpusna livngistika: analiza jezičkih struktura i upotrebe*, Kembridž, Cambridge University Press

[2] Buhler, K., (1934) *Teorija o jeziku: representacijska funkcija jezika*, John Benjamins Publishing Company Amsterdam/Filadelfija

[3] Carter, R., (1993) *Uvod u primenjenu lingvistiku*, Harlou, Penguin

[4] Cook, G., (1990) *Diskurs*, Oksford, Oxford University Press

[5] Crystal, D., (1992) *Uvod u lingvistiku*, Harlou, Penguin

[6] Crystal, D., (1995) *Kembridžova enciklopedija engleskog jezika*, Kembridž, Cambridge University Press

[7] Dakowska, M., (2001) *Psycholingwistyczne podstawy dydaktyki języków obcyc*,. Warszawa, PWN

[8] Dobrić, N., (2008) *Diskurs diskriminacije u engleskom i srpskom*, Filofozski fakultet, Novi Sad, master rad

[9] McCarthy, M., (1991) *Analiza diskursa za nastavnike jezika*, Kembridž, Cambridge University Press

[10] Meyer, C., (2002) *Engleska korpusna lingvistika: uvod*, Kembridž, Cambridge University Press

[11] Stevanović, M. et al., (1967-1976) *Rečnik srpskohrvatskoga književnog jezika Matice srpske*, Novi Sad, Matica srpska

[12] Sapir, E., (1983) *Odabrani spisi Edwarda Sapira o jeziku, kulturi, i ličnosti*, University of California Press

[13] Whorf, B., (1956) *Jezik, misao i realnost: odabrani spisi Benjamina Leea Whorfa*, MIT Press