NIKOLA DOBRIĆ*
Visoka poslovna škola strukovnih studija, Novi Sad

UDK: 81-11

KORPUSNA LINGVISTIKA KAO OSNOVNA PARADIGMA ISTRAŽIVANJA JEZIKA

1. UVOD

Promovisanje termina *korpusna lingvistika* devedesetih godina dvadesetog veka je predstavljalo važnu prekretnicu u procesu prerastanja korpusnog pristupa istraživanju jezika u jednu od preovlađujućih lingvističkih metodologija (Leon 2007: 36). Nalazeći se pod dominacijom generativnog pristupa, on je egzistirao na marginama jezičkog istraživanja. Od pedesetih godina prošlog veka taj pristup lingvistici počeo je da dobija sve veći broj privrženika čime i njegov značaj raste. A u novije vreme on konačno dobija svoje mesto koje mu zasluženo i pripada.

Objavljivanjem niza radova pedesetih i šezdesetih godina prošlog veka Noam Čomski je premestio fokus lingvistike se empirizma ka racionalizmu. Jezička *kompetencija* postaje najvažniji pojam lingvističkog proučavanja. Kritičari korpusnog pristupa jezik u upotrebi smatraju siromašnim ogledalom kompetencije, a korpus siromašnim sredstvom za modeliranje kompetencije. Međutim, kritika koju Čomski iznosi je sa jedne strane usporila razvoj korpusne lingvistike, dok je sa druge strane imala pozitivan uticaj. Kao direktni odgovor na kritiku javila su se mnoga poboljšanja i u teoriji i u praksi korpusnog pristupa (kao na primer reprezentativnost i nepristrasnost izvora) (McEnery i Wilson 2001: 5). Možda zbog pomenutih poboljšanja u narednim decenijama prepoznate su, ili ponovo otkrivene, prednosti istraživanja jezika zasnovanog na autentičnom jeziku. Povećanjem interesovanja za korpusnu lingvistiku osamdesetih godina prošlog veka pojavilo se sve više argumenata u korist korpusnog pristupa.

Naravno, rad sa prirodnim jezikom i empirijski pristup lingvističkoj analizi nije nova stvar i datira mnogo pre pojave prvih modernih korpusnih lingvista. Prvi filolozi su još uvideli, u toj eri empirijskog istraživanja, da su im, pre nego što mogu doneti bilo kakve zaključke o funkcionisanju jezika, potrebni pouzdani jezički podaci. Pronašli su ih u velikim skupovima teksta koji su zamišljeni da predstavljaju reprezentativni uzorak prirodnog jezika. Rad sa skupovima prirodnog jezika je vezan za formulisanje zakonitosti koje proizilaze iz semantičkih obeležja jezika. Jer, dok je sposobnost govora urođena svakom članu jedne društvene zajednice, jezik koji proizvodimo, odnosno značenje koje iskazujemo, je gotovo potpuno društveno uslovljeno. Rezultat toga je da je jezik proizvod proizvoljnog idiosinkretizma i anomalija u jednakoj meri

* Kontakt podaci (Email): goody@sezampro.yu

sa zakonitostima gramatike koje deluju unutar njega. Prirodni jezik u svakodnevnoj upotrebi je sistem u kome je funkcionisanje unutrašnjih elemenata određeno ne zakonitostima spoljnog sveta nego dogovorom oko značenja unutar samog jezika.

Ovakva samoreferentnost jezika i oslanjanje na društveno iskustvo ukazuje na to da empirijski pristup lingvistici koji se ogleda u korpusnom istraživanju ne vuče korene iz opisivanja ontoloških jezičkih univerzalija. Primarna svrha korpusne lingvistike je opisivanje sadržaja i odnosa unutar autentičnog jezika, i izučavanje diskursa kao medijatora znanja i društvenih normi. Korpusni pristup posmatra zakonitosti u jeziku kao uslovljene fundamentalnom praktičnom upotrebom u društvenoj interakciji zajedno sa relevantnim kognitivnim i pragmatičkim implikacijama (Hopper 1998: 156).

Ipak, nasuprot očiglednom značaju i pogodnostima korpusa kao podloge za naučno istraživanje činjenica je da je takav pristup, kako ga vidi korpusna lingvistika, na domaćoj lingvističkoj sceni relativno zapostavljen. Jezičke teorije se često ne verifikuju ispravnim kvantitativnim metodama nego samo sopstvenom introspekcijom, dok relevantnih računarskih korpusa srpskog jezika praktično i nema.

2. NOVA PARADIGMA ISTRAŽIVANJA JEZIKA

Ono što čini ovaj pristup lingvističkom istraživanju toliko značajnim je činjenica da se može postaviti kao nova paradigma istraživanja jezika nasuprot onoj zastupljenoj u *univerzalnoj gramatici* (universal grammar). Pojava koncepata *uslovljene gramatike* (emergent grammar) (Hopper 1998) i *funkcionalne gramatike* (functional grammar) (Halliday 1985) poslužila je za postavljanje primarnih teorijskih podloga empirijskoj analizi jezičkih zakonitosti. Pragmatički pristup gramatici pokazuje da jezičke strukture proizilaze iz upotrebe jezika i da su prvenstveno oblikovane i prethodnim diskursom, odnosno prethodnim jezičkim iskustvom i trenutnim diskursom koji je u toku. Kroz svoju društveno-interaktivnu i kognitivnu dimenziju ovakav pogled na gramatiku kroz korpusni metod posmatra jezik u stanju neprekidne promene. Jezičke strukture su samo delimično ili potpuno okamenjene formacije nastale sedimentacijom često upotrebljavanih oblika čija upotreba zavisi od datog konteksta. Gramatika očigledno nije povlastica pojedinca koju on može upotrebljavati po svom nahođenju, već je gramatika stvar društvenog dogovora unutar jedne diskursne zajednice i zavisi od komunikacijske potrebe u datom trenutku. Zbog toga u ovakvom korpusnom pristupu nema mesta teoretisanju o velikim istinama urođene *jezičke kompetencije*. Jezik je neraskidivo vezan za društvo i jedino se kao takav može i analizirati. Fokus je na *jezičkoj performansi* (*language performance*). Dalje razumevanje i potvrda semantički baziranih teorija o jeziku će sigurno proisteći iz analize velikih skupova teksta, odnosno korpusa (Sinclair 1991: 489).

3. KORPUS

Korpus u svom osnovnom značenju predstavlja skup teksta, bilo pisanog ili govornog jezika. Ipak, posmatrano sa stanovišta korpusne lingvistike, postoji određena

distinkcija unutar samog termina. Skup ili arhiva tekstova koji ne moraju biti podvrgnuti selekciji i ne moraju poštovati određene zahteve niti lingvističke kriterijume se opisuje kao *zbirka tekstova*. Zbirke tekstova, u koje spadaju na primer jedna knjiga ili snimak neke televizijske emisije ne predstavljaju dobru polaznu tačku u istraživanju, jer jezik koji se nalazi unutar tako malog uzorka nikako ne može biti dovoljno reprezentativan. Recimo da se u jednoj analizi istražuje upotreba ličnih zamenica muškog i ženskog roda kao pokazatelja indirektne diskriminacije u jeziku. Za korpus se odabere desetak ili dvadeset knjiga, na primer. Posle izvršene analize sve što se može reći o rezultatima je da je taj broj pisaca na takav i takav način upotrebio date zamenice. Teško se na osnovu takvog nereprezentativnog uzorka može ozbiljno tvrditi da su rezultati do kojih se došlo takvom analizom primenljivi na celokupni jezik. Iako ni korpusi o kojima je reč u ovom radu ne predstavljaju jedan jezik u svojoj celokupnosti, širina i nepristrasnost uzoraka koji čine takve korpuse ipak dozvoljavaju mnogo bolju podlogu za takve generalne tvrdnje.

U korpusnoj lingvistici, kao i u daljem tekstu, termin *korpus* se prvenstveno odnosi na *računarski korpus*. Računarski korpusi su kodirani i standardizovani, optimizovani za pretragu i analizu i nalaze se pothranjeni u računarskim bazama. Obično se sastoje od više miliona reči iz različitih jezičkih i društvenih izvora i idealno obuhvataju sve moguće pojave jednog jezika „uhvaćene" u vremenu i pretočene u elektronski tekstualni oblik.

Postoje korpusi različitih veličina i strukture (*opšti* i *specijalni* korpusi) namenjeni različitim vrstama lingvističke analize. Specijalni korpusi su sakupljeni tako da predstavljaju samo određeni varijetet jezika relevantan za neko polje istraživanja (npr. Korpus kontrole leta (*Air Traffic Control Corpus*)). Opšti korpusi su privlačniji široj paleti naučnih polja kako njihov dizajn predstavlja celokupnu formu jednog jezika u svojoj prirodnoj upotrebi. Najvredniji korpusi opšteg tipa su *monitoring* korpusi. To su korpusi koji održavaju svoju reprezentativnost stalnim dodavanjem novih delova jezika i stalnim proširivanjem varijeteta u njima.

Sledi da je bitno ispravno odabrati i definisati korpus za analizu da bi se osigurala relevantnost povratnih informacija. Prema delu prirodnog jezika koji predstavlja dati korpus postavlja se i opseg i cilj jezičkog istraživanja. Na primer, ako je korpus sastavljen od akademskih tekstova teško se može očekivati da pruži podlogu za analizu varijeteta jezika.

Najvažnija osobina jednog korpusa je njegova reprezentativnost. Reprezentativnost jednog korpusa, a posledično i rezultata koje taj korpus pruža prilikom neke analize, postiže se ne veličinom nego prvenstveno raznolikošću, odnosno pravilnim i planiranim odabirom izvora pri konstrukciji. Korpus mora sadržati što je moguće širi raspon jezičke građe sa uzorcima jednakih veličina, a svi izvori moraju biti precizno deklarisani i obeleženi.

Takođe se mora imati na umu i sinhrona ili dijahrona dimenzija jednog korpusa, gde dijahroni korpusi dodaju dimenziju analize istorijskog razvoja teksta.

Reprezentativnost i ispravno definisanje i odabir korpusa pri analizi je od ključne važnosti da bi se dato istraživanje zaštitilo od toga da korpus slučajno ne predstavlja bilo koju relevantnu promenljivu u dovoljnom stepenu u kome bi se našla u stvarnoj komunikaciji. Takvi se pak nedostaci mogu donekle popraviti primenom raznih modela statističke verovatnoće.

Jedan od prvih i najpoznatijih računarskih korpusa je svakako *Britanski nacionalni korpus* (British National Corpus). To je veliki računarski korpus koji se sastoji od preko 100 miliona reči iz pisanih i govornih izvora. Izvori 75% pisanog jezika su uglavnom informativni tekstovi iz oblasti nauke, religije, ekonomije, filozofije, umetnosti i medija, dok je 25% odvojeno za književna dela. Usmeni jezik je zastupljen u oko 10 miliona reči i sastavljen je od transkripta spontanih razgovora, skriptovanih razgovora, javnih govora i usmenog jezika u medijima. Korpus je etiketiran za vrste reči i većina frekvencijskih pretraga se ne naplaćuje. Zbog svoje veličine i reprezentativnosti Britanski nacionalni korpus se smatra najboljim postojećim korpusom britanskog engleskog jezika i može se pronaći na http://www.natcorp.ox.ac.uk/.

Među slične korpuse spadaju i:

- Korpus savremenog američkog engleskog jezika (*Corpus of Contemporary American English):* monitoring korpus opšteg tipa sa 360 miliona reči na adresi http:// www.americancorpus.org/;
- Korpus australijskog engleskog jezika (*Australian Corpus of English*): monitoring korpus opšteg tipa sa 1 milionom reči na adresi http://khnt.hit.uib.no/icame;
- Kembridžov međunarodni korpus (*Cambridge International Corpus*): višejezični korpus specijalnog tipa sa 275 miliona reči na adresi http://www.cambridge.org/elt/corpus;
- Ruski nacionalni korpus (*Russian National Corpus*): monitoring korpus opšteg tipa sa 150 miliona reči na adresi http://www.ruscorpora.ru/en/index.html;
- Nacionalni korpus hrvatskog jezika: monitoring korpus opšteg tipa sa 30 miliona reči na adresi http://www.hnk.ffzg.hr/;
- Korpus savremenog srpskog jezika: korpus opšteg tipa sa 24 miliona reči na adresi http://korpus.matf.bg.ac.yu/prezentacija/korpus.html;
- Korpus srpskog jezika: korpus opšteg tipa sa 12 miliona reči na adresi http://www.serbian-corpus.edu.rs/indexns.htm.

Drugi poznati korpusi su: Korpus starogrčkog jezika (*Thesaurus Linguae Graecae*), Francuska međujezička baza (*French Interlanguage Database*), Lankasterski korpus mandarinskog kineskog (*Lancaster Corpus of Mandarin Chinese*), Korpus dečjeg jezika (*Child Language Data Exchange System*), Vulverhempton korpus engleskog poslovnog jezika (*Wolverhampton Business English Corpus*), Korpus srednjevekovnog engleskog stiha i proze (*Corpus of Middle English Prose and Verse*), itd.

4. TERMINOLOGIJA

Svrha ovog poglavlja je da predstavi neke od osnovnih termina korpusne lingvistike. Termini će biti predstavljeni u izvornom obliku na engleskom jeziku, a potom prevedeni i dodatno objašnjeni:

coding – kodiranje. Postupak dodavanja dodatnih lingvističkih informacija u tekstove unutar korpusa;

tagged/untagged corpora – etiketirani/neetiketirani korpusi. Etiketiranje[1] je komplikovani proces dodavanja dodatnih informacija u korpus. Informacije se mogu ticati obeležavanja reči po gramatičkoj kategoriji, rodu, broju, morfološkim i fonološkim karakteristikama, itd. Etiketirani korpusi su oni koji sadrže neke od pomenutih dodatnih informacija a neetiketirani sadrže samo tekst bez dodatnih lingvističkih podataka;

parsing – parsiranje. Postupak dovođenja morfosintaksičkih kategorije u tekstu u međusobne sintaksičke odnose višeg stepena. Parsiranje je postupak odvajanja rečeničnih delova i opisivanje odnosa između njih. Parsiranjem se određuje sintaksička struktura rečenice i retki su korpusi koji poseduju ovakav napredni nivo kodiranja;

type vs. *token* – tip i token. *Tip* se odnosi na kvalitativnu analizu i predstavlja istraživanje vezano za određene tipove reči (npr. određene vrste reči ili određene gramatičke kategorije) ili lekseme, dok se *token* odnosi na pojedinačno javljanje neke reči čija frekventnost spada u polje kvantitativne analize;

raw frequency count; *operation* – sirova frekventnost; operacija. Oba termina znače isto a odnose se na neobrađene podatke o brojnosti analizirane lingvističke jedinice u korpusu. Na te brojčane vrednosti frekventnosti se dalje primenjuju razne matematičke i statističke funkcije u cilju dobijanja podataka na osnovu kojih se mogu izvlačiti određeni zaključci;

concordancer – konkordanser. Da bi se došlo do rezultata frekventnosti pri analizi nekog korpusa koriste se posebni računarski programi koji se nazivaju *konkordanseri*. Konkordanseri dozvoljavaju razne vrste pretraga korpusa, od jednostavne analize frekventnosti, preko istraživanja po vrsti reči ili morfološkim nastavcima, istraživanja kolokacija pa sve do fonoloških pretraga teksta. Pored svoje korisnosti, jedna od prednosti konkordansera je i ta što su veoma jednostavni za upotrebu.[2] Svi korpusi koji se mogu pronaći na Internetu dolaze sa sopstvenim ugrađenim konkordanserima.

5. OSNOVNE KARAKTERISTIKE KORPUSNOG PRISTUPA

Korpusni pristup je primenljiv na više različitih polja, koje čine samo jezgro lingvistike kao nauke:

- leksikografija (upotreba korpusa je veoma rasprostranjena pri pravljenju rečnika);
- sociolingvistika (korpusni pristup omogućuje istraživanje dijalekata, registara i samog društva);
- analiza diskursa (ovakav pristup obezbeđuje dovoljno velike uzorke diskursa omogućujući time pronalaženje karakteristika jezika bez strukturalnih ograničenja);
- morfologija (rezultati dobijeni pri analizi korpusa mogu otkriti mnogo o frekventnosti, distribuciji i ulozi raznih oblika leksema);
- fonologija (računarski korpusi mogu pružiti uvid u različite pojave fonetske distribucije i pomoći u pronalaženju zakonitosti);

- semantika (teško je pronaći pristup koji može pružiti toliko podataka o značenju reči kao korpusni pristup);
- sintaksa (istraživanje jezičkih struktura na ovakav način može pružiti empirijske dokaze o tome kako pristupamo konstruisanju rečenica i kako se izražavamo kroz jezik);
- komparativna i kontrastivna lingvistika (postojanje paralelnih korpusa može otkriti sličnosti i razlike među jezicima);
- metodika nastave (korpusi mogu pomoći pri dizajniranju materijala i aktivnosti za učenje jezika);
- kognitivna lingvistika (autentična prirodna upotreba jezika smeštena u korpuse daje uvid u način na koji mentalni procesi utiču na komunikaciju i na jezik u celini).

Zajedničko svakoj analizi koja bi potekla iz bilo koje od nabrojanih lingvističkih disciplina je to da je prilaz istraživanju najčešće induktivnog, a ne uobičajenog hipotetičko–deduktivnog karaktera. To znači da su katalizatori istraživanja konkretni podaci koji usmeravaju teorijske zaključke (*data–driven* pristup) umesto prethodno zamišljenih pravila i teorija.

Korpusni pristup analizi jezika ima nekoliko osnovnih odlika koje omogućavaju skoro svaku vrstu lingvističke analize, a svakog lingvistu čine potencijalnim korisnikom korpusa. U pitanju su sledeće odlike korpusnog pristupa:

- empirijskog je karaktera budući da se bavi analizom komunikacije u njenom prirodnom obliku;
- analiza se zasniva na velikim skupovima teksta koji predstavlja jezik, a koji se nazivaju korpusi;
- koriste se računari u istraživanju;
- fokus je na jezičkoj *performansi* umesto na jezičkoj *kompetenciji*;
- radi se o *kvantitativnom* i *kvalitativnom* modelu proučavanja jezika (Leech 1992: 106-107).

Svaka od ovih osobina čini korpusni pristup lingvistici veoma atraktivnim i privlačnijim od drugih pristupa istraživanju jezika. Činjenica da je empirijskog karaktera ukazuje na to da se radi ne o veštački stvorenom jezičkom skupu nego o prirodnom tekstu. S obzirom na to da je reč o veoma velikim i reprezentativnim skupovima autentičnog jezika moguće je zamisliti koliko opcija ovakav pristup pruža u pravcu mogućih tema otvorenih za istraživanje. Dalje, upotreba računara i računarskih programa pruža lingvistima mogućnost da izađu na kraj sa veoma kompleksnim analizama velikih baza podataka pružajući visok stepen tačnosti rezultata analize.

6. JEDINICE ANALIZE

Kao što je već pomenuto, računarski korpusi se mogu koristiti kod različitih vrsta istraživanja jezika. Jezičke jedinice analize se razlikuju od istraživanja do istraživanja, od manjih segmenata u fonologiji i morfologiji do složenijih jedinica

u sintaksičkoj i posebno u semantičkoj analizi. Upravo jedinice semantičke analize zbog svoje kompleksnosti zahtevaju detaljnije definisanje sa stanovišta korpusnog istraživanja.

Iz perspektive korpusne lingvistike leksičke jedinice mogu biti pojedinačne reči ili složenije značenjske jedinice. One su u principu *monosemne*. To je ono što razlikuje pojam leksičke jedinice od pojma reč. Većina reči je zapravo *polisemna*. Složene leksičke jedinice se mogu posmatrati kao osnovna reč zajedno sa svim rečima u kontekstu sa kojima formira jednu semantičku jedinicu (Teubert 2005:4). Sve dok je značenje takve jedinice nejasno ona nije potpuna. Potpuna je samo kada predstavlja jedno jedinstveno značenje kao jedna jezička celina bez obzira na broj reči.

7. KVANTITATIVNA I KVALITATIVNA ANALIZA

Korpusna lingvistika poseduje još jednu prednost. Zakonitosti i zaključci koji proizilaze iz istraživanja korpusa zasnivaju i na *kvalitativnoj* i na *kvantitativnoj* analizi. Kvalitativna dimenzija istraživanja se odnosi na istraživanje jezičkih pojava kao određenih tipova, predstavnika svoje jezičke grupe (npr. istraživanje imenica, ili određenih morfoloških nastavaka, kao predstavnika jedne klase) dok se kvantitativna analiza odnosi na frekventnost pojedinačnih jezičkih jedinica, odnosno *tokena*. Kvalitativna analiza uveliko zavisi od nivoa etiketiranosti korpusa i prilično je težak zadatak ako se izvodi na velikim skupovima teksta, tako da je rezervisana prvenstveno za korpuse manjeg obima. Kvalitativna analiza se često koristi kao katalizator prvih teorija koje se zatim ispituju na većem uzorku jezika mnogo robusnijom metodom kvantitativne analize. Kvantitativna analiza je fokusirana prvenstveno na frekventnost tokena i služi da se kroz upotrebu pojedinačnih jedinica provere teorije postavljene za ponašanje prethodno kvalitativno ispitanih tipova.

8. FREKVENTNOST I STATISTIČKA ANALIZA

Osnovni parametar zakonitosti unutar jednog korpusa jeste frekventnost referentnih jezičkih jedinica (fonema, morfema, leksema, semema) uzetih kao sadržaj analize. Frekventnost označava koliko se puta neka jezička jedinica pojavljuje u određenom korpusnom kontekstu, što posle dozvoljava konstruisanje određenih zaključaka vezanih za temu istraživanja.

Brojčane vrednosti frekventnosti služe kao podloga za ono što čini samu srž korpusne lingvistike: statističku analizu. Razne vrste statističke obrade podataka su zapravo materija koja daje čvrstu empirijsku bazu i služi kao izvor svakoj posledičnoj teoriji o nekoj jezičkoj pojavi. Statistički proračuni se koriste i kod izračunavanja verovatnoće pojave jezičke jedinice u pretpostavljenom diskursu teorijski neograničene veličine. Statistička analiza je obavezan deo svakog korpusnog istraživanja jer ona ne samo da daje temelj teoretisanju o određenim jezičkim zakonitostima i izračunavanju verovatnoće, nego i potvrđuje naučnu vrednost podataka verifikujući ih ili kao nasumične ili kao lingvistički relevantne (Johnson 2008:5).

Čini se nedopustivo da i danas postoje razne teorije i tvrdnje iznete na osnovu analize datog korpusa bez prethodne statističke verifikacije polaznih podataka. Da bi se kategorički moglo tvrditi da rezultati dobijeni iz jednog korpusa nisu samo puka slučajnost i da bi se dobijene vrednosti sa uzorka mogle generalno proširiti na jezik u celini potrebno ih je matematički obraditi i statistički potvrditi. Razlog za izbegavanje statističkih metoda u lingvistici je verovatno taj što sve te matematičke formule izgledaju strano većini lingvista pa ih većina i izbegavaju. Međutim, izbegavati takve naučne metode nije ispravno jer inače dobijeni rezultati ostaju i dalje u teorijskom nedokazanom obliku. Izbegavanje statističkih metoda je dalje neopravdano jer pored mogućnosti okretanja profesionalnoj pomoći (statističarima), postoje i razna softverska rešenja koja olakšavaju ovakve nepopopularne procedure.

9. KONTEKSTUALIZACIJA

Pored rada sa autentičnim jezikom i pored činjenice da većina prihvatljivih računarskih korpusa, osim diskursne reprezentativnosti, obezbeđuju etiketiranjem i dodatne informacije o jezičkim jedinicama, korpusni pristup poseduje još jednu prednost, a to je kontekst. Kontekst igra jednu od presudnih uloga u odabiru gramatičkih rešenja koje ćemo koristiti u datoj komunikacionoj situaciji. Jasno je da su onda informacije o kontekstu i okolnostima komunikacije od neizmerne važnosti. Svaki korpus pruža precizno definisani uvid u kontekstualnu situaciju u kojoj je dati tekst proizveden, što nam daje kontrolu nad još jednom varijablom u istraživanju dajući nam na taj način više kontrole nad istraživanjem.

10. VRSTE REZULTATA

Korpusna lingvistika daje opšte i pojedinačne teorije o diskursu bazirane na analizi odgovarajućeg preseka diskursa, što zapravo korpus i predstavlja. Opšte zakonitosti su vezane za pravila ili statističku verovatnoću određenih pojava i spadaju uglavnom u domen istraživanja gramatike kao funkcije semantike ako se analizirana jezička jedinica može smatrati tipom. Pojedinačne zakonitosti se tiču tumačenja datih delova teksta ne kao predstavnika svojih klasa nego kao jedinstvenih komunikacijskih pojava, odnosno tokena. Pribavljeni rezultati uvek za cilj imaju teoretisanje na osnovu sistematično klasifikovanih lingvističkih elemenata iz date korpusne analize kroz niz deduktivnih procesa. Tako dobijene teorije uvek nastaju kao opisi jezičke performanse i zakonitosti upotrebe, značenja jezika ili relevantnih kognitivnih procesa.

11. OGRANIČENJA KORPUSNOG PRISTUPA

Najveći nedostatak korpusnog pristupa lingvistici može se sumirati u sledećoj izjavi: ,,Mislim da ne postoji takav korpus, ma koliko bio veliki, koji bi posedovao dovoljno podataka o svim oblastima leksikona i gramatike [engleskog] jezika koje bih ja

želeo da analiziram [...]."[3] Tačno je da zamerke koje stižu na račun korpusne lingvistike, koje tvrde kako bilo koji obim autentičnog jezika i komunikacije nikada neće uspeti da predstavi apsolutno sve oblike jezika imaju određenu težinu, jer se ni svi teoretski postojeći oblici ne pojavljuju u stvarnoj komunikaciji zbog kontekstualnih ograničenja i rutinske dimenzije upotrebe jezika. Potpuno je tačno da i najreprezentativniji računarski korpusi koji postoje predstavljaju samo jedan, veći ili manji, presek apsolutnog diskursa, i vrlo je verovatno da ni stvaran jezik i stvarna jezička izvedba zapravo i ne proizvodi sve zamišljene oblike jezika. Da i pored ovakvih teorijskih nedostataka, koji se ne uklapaju u učenje univerzalne gramatike, korpusni pristup zaslužuje centralno mesto u proučavanju jezika potvrđuje nastavak prethodno spomenute izjave: „[...] ali svaki korpus koji sam imao prilika da analiziram, bez obzira na to koliko je mali bio, prikazao mi je činjenice koje ne bih ni na kakav drugi zamisliv način mogao pronaći."[4]

12. KORPUSNA LINGVISTIKA I RAČUNARSKI KORPUSI U SRBIJI

Pored detaljno opisanih uslova koji čine računarske korpuse prihvatljivim u empirijskom naučnom istraživanju najvažniji uslov je naravno postojanje samog reprezentativnog računarskog korpusa. Može se činiti iz svega što je navedeno kako se takav uslov mora uzeti kao dat jer je postojanje jedne tako bitne stvari sigurno zagarantovano. Nažalost to nije slučaj sa srpskim jezikom (Dobrić 2009), iako rad na računarskim korpusima ima dugu istoriju na ovim prostorima. Prvi korpus srpskog (srpsko-hrvatskog) jezika napravljen je samo godinu dana posle prvog većeg računarskog korpusa u svetu.[5] Rad na korpusima i učestvovanje na velikim međunarodnim projektima je trajao sve do devedesetih godina dvadesetog veka, kada je saradnja prekinuta. U poslednjih desetak godina bilo je pokušaja da se saradnja nastavi, ali u nedovoljnom obimu. Takođe je i primetan nedostatak opšteg nacionalnog korpusa srpskog jezika.[6]

Trenutno postoje samo dva računarska korpusa srpskog jezika (Korpus savremenog srpskog jezika i Korpus srpskog jezika) koji nažalost ne ispunjavaju potrebne kriterijume. Nedostaci datih korpusa se tiču prvenstveno reprezentativnosti, jer prvo ni jedan od njih nema u izvore uvršten govorni jezik, što je velika mana. Dalje, *Korpus savremenog srpskog jezika*, koji je proizišao iz projekta *Matematička i računarska lingvistika* 1981. godine pod vođstvom Duška Vitasa, i posle postavljanja na internet većim delom ostaje neetiketiran. Izvori za 24 miliona reči, od kojih dve trećine čine tekstovi iz Politike, nisu ni približno dovoljno raznovrsni. Ni *Korpus srpskog jezika*, vizionarski započet još 1957. od strane Đorđa Kostića, kao deo velikog jezičkog projekta socijalističke Jugoslavije na kome su učestvovali i Rudolf Filipović i Željko Bujas (Tadić 1997: 388), ne zadovoljava kriterijume reprezentativnosti. Korpus je pretvoren u elektronski tekst 1996. od strane Aleksandra Kostića, koji je nastavio očev rad, i sadrži 11 miliona reči. Korpus poseduje odličnu dijahronu dimenziju sa izvorima počevši od 12. veka. Nedostaci se tiču sinhrone dimenzije jezika koja praktično ne postoji, jer nedostaju uzorci savremenog srpskog jezika[7].

Zbog opisanih nedostataka postojeći korpusi nisu dovoljno pogodni da bi mogli pretpostaviti da predstavljaju celokupni diskurs srpskog društva, što je ciljni oblik koji

je od najveće koristi društvenim naukama. Očigledno je da srpskoj naučnoj zajednici nedostaje višemilionski nacionalni korpus (Tadić, 1990) koji bi služio kao riznica srpskog jezika i koristio sveukupnoj paleti naučnih disciplina koje mogu svoje istraživanje bazirati na jeziku. Neophodno je sastavljanje novog, sveobuhvatnijeg nacionalnog korpusa srpskog jezika, 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.

1 Proces se naziva i anotiranje (Kostić 2003: 260).
2 Za više informacija o konkordanserima pogledati Bol (Ball 1996).
3 "I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore [...]". (Fillmore 1992: 35)
4 "[...] but every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way." (Fillmore 1992: 35)
5 Brown korpus (http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html)
6 Autor je više pisao o istoriji korpusnog pristupa u Srbiji u preglednom članku koji će uskoro biti objavljen.
7 Važno je napomenuti da je Korpus srpskog jezika potpuno ručno etiketiran za gramatički status, broj grafema, slogova, za fonološku strukturu i za početak i kraj rečenica. Ovakav nivo etiketiranja je veoma redak i čini ovaj korpus jednim od najvećih ručno obrađenih korpusa u svetu (Kostić 2003: 260).

LITERATURA

Ball, C. 1996. *Tutorial Notes: Concordances and Corpora*. Washington DC: Georgetown University.

Dobrić, N. 2009. *Korpusna statistika u lingvistici – Mogućnost primene u izučavanju jezika*. Predavanje u Statističkom društvu Srbije – Klub statističara Vojvodine, Novi Sad, štampani tekst predavanja.

Fillmore, C. J. 1992. Corpus Linguistics or Computer-Aided Armchair Linguistics. In J. Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* 82, 35-60.

Haliday, M. A. K. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Hopper, P. 1998. Emergent Grammar. In M. Tomaselo (ed.) *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, 67-92.

Johnson, K. 2008. *Quantitative Methods in Linguistics*. Oxford: Blackwell Publishing.

Kostić, A. Đ. 2003. Elektronski korpus srpskog jezika Đorđa Kostića. *Zbornik Matice srpske za slavistiku* 64, 260-264.

Leech, G. 1991. The State of the Art in Corpus Linguistics. In K. Aijmer & B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, 8-29.

Leech, G. 1993. Corpus Annotation Schemes. *Literary and Linguistic Computing* 4, 275-281.

Leon, J. 2007. Claimed and Unclaimed Sources of Corpus Linguistics. *Henry Sweet Society Bulletin* 44, 36-50.

McEnery, T. & A. Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Meyer, C. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
Sinclair, J. 1991. *Corpus, Concordance, Collocation: Describing English Language*. Oxford: Oxford University Press.
Tadić, M. 1990. Zašto nam je potreban višemilijunski referentni korpus? *Informatička teorija u primenjenoj lingvistici*, 95-98.
Tadić, M. 1997. Računalna obradba hrvatskih korpusa: povjest, stanje, perspektive. *Suvremena lingvistika* 43-44, 387-394.

SUMMARY

CORPUS LINGUISTICS AS THE BASIC PARADIGM OF LANGUAGE RESEARCH

The introspection–based analysis has been the prevailing linguistic paradigm for a long time. The corpus approach in its contemporary framework marks the return of linguistics within the boundaries of empirically founded sciences, where a scientific discipline which studies language surely does belong. This paper first elaborates on the advantages of such a research approach which focuses on the language community using a given language and the nuances of their usage. The paper goes on to describe the basic characteristics of the corpus approach to linguistics. In the end the paper stresses the importance of constructing a general representative computer corpus of the Serbian language.

KLJUČNE REČI: korpus, diskurs, frekventnost, token, konkordanser, statistika.