UDK 81'322
81'11:004

TANJA SAMARDŽIĆ
Filološki fakultet u Beogradu

15

OSNOVNE TEORIJSKO-METODOLOŠKE KARAKTERISTIKE KORPUSNE LINGVISTIKE

The language looks rather different
when you look at a lot of it at once.
(Sinclair 1991: 100)

Od trenutka kada je konstruisan prvi kompjuter postoji veliki interes da se jezik opiše (ili obradi) na način koji bi omogućio da se ovaj izum primeni kao sredstvo za automatizaciju u različitim sferama jezičke produkcije. Istraživanja u ovoj oblasti karakterišu se različitim ciljevima, pristupima i tehnikama. Iako rezultati, uglavnom, ne odgovaraju uloženim naporima, kada je u pitanju primena softverskih proizvoda, problemi na koje se u tim istraživanjima nailazi predstavljaju značajan izvor podataka o različitim jezičkim pojavama. U poslednjih desetak godina, međutim, može se primetiti izdvajanje jedne posebne grane lingvistike čiji se metodi, koncepti i teorijske postavke zasnivaju na obilnom korišćenju kompjuterskih resursa u *analizi* jezika. Danas je uobičajeno da se ova oblast naziva **korpusna lingvistika**. U današnjem shvatanju pojma korpusne lingvistike presudnu ulogu ima projekat *Cobuild* koji je 1980. godine započet na Univerzitetu u Birmingamu, u saradnji sa izdavačkom kućom *HarperCollins*. Većina ideja koje predstavljamo u ovom tekstu potiče iz ovog projekta.

Karakteristično za ovaj novi pogled na jezik jeste to da se u centru interesovanja nalazi **tekst** kao realizovan jezik, kao ono što se realno događa u jeziku, sa svim svojim neregularnostima, posebnostima, idiosinkretičnošću, asimetrijom, izuzecima, disproporcijom i svim onim što oduvek muči lingviste i zbog čega je i sam De Sosir na izvestan način izdvojio ovaj aspekt jezika smatrajući da on ne treba da predstavlja predmet lingvističkih istraživanja. Savremeni memorijski kapaciteti kompjutera omogućavaju da se skladištene velike kolekcije tekstova (**korpusi**) obraduju na različite načine. Iz korpusa se automatski ekscerpiraju ogromne količine različitih podataka o ponašanju jezičkih jedinica, koji se zatim

interpretiraju pomoću različitih manje ili više komplikovanih statističkih tehnika (procenti, verovatnoća, faktorijalna analiza) (v. Biber 1988), što je eksplicirano i u jednoj od novijih reprezentativnih studija.

> "Ovaj pristup predstavlja ono što danas ide pod imenom korpusna lingvistika: način ispitivanja jezika posmatranjem velikih količina spontano nastalog, elektronski skladištenog diskursa, uz korišćenje softvera koji odabira, sortira, poredi, broji i računa." (Hunston 1999: 15)

Korpus se uzima kao uzorak jezika nad kojim se vrše analize, merenja i eksperimenti, da bi se induktivnim putem došlo do uopštavanja i zaključaka. Teži se tome da analiza bude što je moguće objektivnija i da se što manje oslanja na uspostavljene tradicionalne lingvističke kategorije. Karakterišući jezičke podatke koji se analiziraju u korpusnoj lingvistici, Hanston (Hunston 1999: 15) ističe pet aspekata u kojima se oni razlikuju od podataka koji se koriste u drugim lingvističkim pristupima. Podaci su 1) autentični, 2) odabrani metodom slučajnog uzorka, a ne prema unapred postavljenim lingvističkim kriterijumima, 3) brojni, 4) sistematski organizovani i 5) nestrukurirani prema bilo kojoj od postojećih teorija.

Ovakav pristup rezultirao je nekim novinama u jezičkoj analizi. Odustaje se od strogih gramatičkih generalizacija. Uopštavanja i kategorizacije se radije izražavaju kao statističke tendencije u okviru **kontinualnih skala**. U tom smislu, karakterističan je stav D. Bajbera (Biber 1988: 22), koji smatra da nema osnova da se očekuje da bi dimenzije variranja u jeziku trebalo da budu dihotomne i da izrazito dihotomna priroda postojećih dimenzija taksonomija samo pokazuje "koliko je preliminarna faza u kojoj radimo." U svojoj studiji, zato, Bajber dimenzije jezičkog variranja identifikuje kao "kontinualne kvantifikabilne parametre varijacije, tj. kontinualne skale."

Korpusna lingvistika naročito ističe značaj **konteksta** u kom se određena jezička jedinica realizuje u tekstu. Značenje jezičkih jedinica sada se u tom smislu tretira kao dinamična kategorija (Daniellsson 2001: 85), izuzetno zavisna od konkretnih tekstuelnih realizacija, tako da se uspostavlja novi odnos između forme i značenja jezičkih jedinica, uz praktično brisanje razlike između tradicionalno odvojenih sfera sintakse i leksikona. To znači da *značenje* jedne jedinice utiče na to kakav će se **obrazac** formirati oko nje u tekstu, ali i obrnuto: da prema utvrđenom obrascu možemo da zaključimo o kakvom se značenju radi. Ovo se posebno odražava na razlikovanje pojedinih značenja višeznačnih leksema. Oslanjajući se na korpusne studije, Sinkler (Sinclair 1991:6–7) tvrdi da se svako značenje jedne jedinice može povezati sa različitim formalnim obrascem koji se formira oko nje. Ovaj autor, dalje, tvrdi da je formiranje ovakvih obrazaca u toj meri regularno, da se može očekivati da će se formalni obrasci u budućnosti otvoreno koristiti kao kriterijumi za analizu značenja. To bi bilo sigurnije i manje ekscentrično polazište za "disciplinu koja pretenduje na naučnu ozbiljnost." Drugim rečima: "Još jednom, stisak intuicije je oslabljen" (Sinclair 1991: 7).

Za ovakvo tretiranje značenja veoma je bitna činjenica da se izvesne jedinice u tekstu u većoj ili manjoj meri **ponavljaju**. Čak i ukoliko nije moguće utvrditi zašto se neke jedinice ponavljaju i zašto su neke pojave u tekstu češće od nekih drugih, neosporno je da ponavljanja na određeni način utiču na status i prirodu jezičkih

jedinica. Zbog toga korpusna lingvistika uzima **frekvenciju** jezičkih jedinica u tekstu kao njihovo značajno svojstvo. Prema Sinkleru (Sinclair 1991: 101), frekventne reči, uopšteno gledano, imaju složeniji skup značenja nego nefrekventne. Akumulacija pojavljivanja neke frekventne reči ne predstavlja samo dodavanje istog, već sve jasniji dokaz kompleksnosti. Ispitivanja u ovom smeru trebalo bi da otkriju izvesne statističke relacije izmedu broja pojavljivanja jedne reči i broja značenja koja ona realizuje.

Osim činjenice da se jezičke jedinice u tekstu ponavljaju, veoma je bitno i to da se neke jedinice često javljaju zajedno sa drugim jedinicama, što se takode smatra kao podatak od značaja u pokušaju da se utvrdi značenje i jednih i drugih. Stabs (Stubbs 2001: 16) smatra da jedan od glavnih dokaza značenja neke reči predstavljaju upravo reči koje se pojavljuju oko nje, posebno one koje se ponavaljaju u odredenim obrascima.

Vezana upotreba odredenih jezičkih jedinica primećena je i ranije u lingvističkim istraživanjima, ali se uglavnom tretirala kao neregularna periferna pojava, bilo kao **idiom, frazema** ili **kolokacija**. U korpusnoj lingvistici, medutim, ovaj aspekt jezičke upotrebe dobija centralno, zapravo *principijelno* mesto, koje Sinkler (Sinclair 1991: 110) definiše kao **princip idioma**. Ovaj princip podrazumeva da korisnik jezika formuliše i razumeva iskaze na osnovu većeg broja "semiprekonstruisanih", tj. polugotovih, fraza koje su skladištene u njegovoj memoriji. Drugim rečima, kad god hoće nešto da saopšti, govornik prvo proveri inventar polugotovih fraza, kako bi našao odgovarajuću. Isto tako, kod razumevanja, svaki primljeni iskaz prvo se poredi sa inventarom fraza, kako bi se pronašlo odgovarajuće značenje. Ukoliko se odgovarajuća fraza ne nalazi u inventaru, govornik onda prelazi na **princip izbora**. Prema ovom principu, od jezičkih elemenata i prema jezičkim pravilima formiraju se sasvim novi iskazi. Kod svakog završetka jedne jedinice (reči, fraze ili rečenice), otvara se veliki spektar mogućnosti izbora sledeće jedinice. Ove mogućnosti su odredene gramatičkim pravilima. Princip idioma ima prvenstvo, tj. jezičke poruke se normalno dekodiraju pomoću principa idioma, prelazak na princip izbora nužno je signaliziran (intonacijom, specijalnim redom reči ili sličnim sredstvima) u tekstu.

Problemi u korišćenju statističkih tehnika (na primer, iskakanje najfrekventnijih reči iz opštih statističkih pravilnosti), kao i usvajanje principa idioma problematizuju i pitanje **jezičkih jedinica**.

"Ljudi govore utvrdenim frazama, pre nego pojedinačnim rečima (...)". (Mel'cuk 1998: 24)

U svom istraživanju Danijelson (Daniellsson 2001) pokazuje da je moguće automatski izdvojiti neke fraze koje funkcionišu kao jezičke jedinice, što potvrduje i činjenica da neki statistički proračuni (Zipfov zakon, na primer) funkcionišu pravilnije kad se primene na pomenute automatski ekstrahovane segmente teksta koji sadrže više od jedne reči nego kad se primene na pojedinačne reči.

Rezultati istraživanja u oblasti korpusne lingvistike paralelno se primenjuju u brojnim komercijalnim izdanjima rečnika i gramatika engleskog jezika kuće *HarperCollins* (v. npr. Combley *et al.* 2002, Francis *et al.* 1996) kao i u daljim

istraživanjima mogućnosti automatizacije procesa izrade rečnika i tezaurusa (Barnbrook 2002). Najnovija istraživanja tiču se primene navedenih principa u oblasti dvojezičnih odnosno višejezičnih rečnika i prevođenja, uz formiranje paralelizovanog višejezičnog korpusa.

LITERATURA

Barnbrook, G. 2002. *Language of Definitions*. Amsterdam: John Benjamins Publishing Company.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Combley, R. *et al.* 2002. *Collins COBUILD New Student's Dictionary*. Glasgow: HarperCollins.

Daniellsson, P. 2001. *The Automatic Identification of Meaningful Units in Language*. Sprakdata, Department of Swedish, Göteborg University (PhD Dissertation).

Francis, G. *et al.* 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.

Hunston, S. and G. Francis. 1999. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins Publishing Company.

Mel'cuk, I. 1998. Collocations and Lexical Functions. In A. P. Cowie (ed.) *Phraseology, Theory, Analysis, and Applications*. Oxford: Oxford University Press, 23–55.

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

Stubbs, M. 2001. *Words and Phrases – Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

SUMMARY

MAIN THEORETICAL AND METHODOLOGICAL FEATURES OF CORPUS LINGUISTICS

This paper reviews the key notions within the corpus linguistics approach. Texts are considered to be scientifically observable, quantifiable language units. Corpus is a sample of language. Continuous scales are seen as a possible instrument for the quantification of linguistic data. Context plays a significant role in the identification and description of linguistic units. Reoccurrence, co-occurrence and frequency are seen as possible evidence for linguistic facts. A pattern is a generalized linguistic unit based on this evidence. The idiom principle is seen as primary when denoting strategy and the open choice principle as secondary.