Pregledni znanstveni rad Primljeno: 17. XI. 2008. Prihvaćeno: 18. XII. 2008.
UDK: 81'42:81'374

Obol korpusne lingvistike suvremenoj leksikografiji

Iva Klobučar Srbić
Leksikografski zavod Miroslav Krleža, Zagreb

SAŽETAK: Korpusna lingvistika zauzela je nezamjenjivo mjesto u suvremenoj leksikografskoj praksi. U ovome radu riječ je o korpusima, posebice računalnim korpusima, jezičnim tehnologijama i alatima, te o primjeni korpusne lingvistike, osobito u suvremenoj leksikografiji. Posebno je poglavlje posvećeno korpusnoj lingvistici u Hrvatskoj, njezinim početcima, ali i suvremenim projektima i postignućima.

Ključne riječi: *korpus, korpusna lingvistika, jezične tehnologije, korpusna leksikografija*

Uvod

Suvremena korpusna lingvistika zasebna je grana lingvistike koja se bavi jezičnom analizom strojno izrađenih korpusa pisanoga ili govornoga jezika. Poput kognitivne lingvistike pretpostavlja empirijski pristup jeziku i zapravo se suprotstavlja chomskyjevskomu gledištu prema kojem se jezična djelatnost odvija u zasebnome jezičnome modulu, neovisno o ostalim mentalnim aktivnostima. Kognitivna lingvistika pak jezičnu djelatnost vidi kao dio ukupne mentalne strukture. Iz toga proizlazi da u razmatranjima o prirodi i funkcioniranju jezika moramo uzeti u obzir njegovu uporabu, jer je upravo uporaba jezika temelj svakoga korpusnoga pristupa jeziku.

Korpusi / Računalni korpusi

Postoji više definicija korpusa, ali najopćenitijom se smatra ona da je korpus zbirka tekstova prirodnoga jezika sastavljena po određenome kriteriju. Pritom treba imati na umu da i zbirku tekstova čine tekstovi skupljeni prema nekim kriterijima te da nije svaka zbirka tekstova korpus. Dakle, korpus je skup jezičnih odsječaka (ne mora biti sastavljen od cijelih tekstova) koji su odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak[1]. Uz

[1] Tadić 2003.

korpus se uvodi i popis svih izvora koji mora biti javan i koji definira tipove korpusa. *Opći korpusi* (referentni) reprezentativni su za jezik u cjelini, a služe za istraživanje raznolikosti na svim jezičnim razinama. Najčešće se nazivaju i *nacionalnim korpusima*[2]. *Specijalizirani korpusi* obuhvaćaju samo jedan jezični varijetet odabran po određenim kriterijima. Rezultati koji se mogu dobiti istraživanjem korpusa jesu:

1. evidencija (nalazi li se neki jezični entitet u tom korpusu ili ne),
2. frekvencija (ako neki jezični entitet postoji u tom korpusu, koliko se puta pojavljuje)[3] i
3. relacija (u kakvom odnosu taj jezični entitet stoji prema svojoj okolini)[4].

Veličina korpusa mjeri se brojem pojavnica (svih riječi koje su se pojavile u korpusu), a ovisi o njegovoj svrsi, ali i o tome da bude što šire primjenjiv. Odluke o veličini korpusa, odnosno kriteriji odabira tekstova koji će ući u korpus (tip, starost teksta, žanr, dob i spol autora, vrijeme nastanka ili objavljivanja), arbitrarne su i ovise o sastavljaču korpusa. Dosadašnja je praksa pokazala da se u svakom korpusu polovica svih različnica (različitih riječi) pojavljuje samo jednom te da nijedan korpus nije tako velik da bi mogao obuhvatiti sve riječi nekoga jezika. Leksik svakoga jezika beskonačan je, otvoren popis u koji nove jedinice svakodnevno ulaze, neke postupno izlaze iz uporabe, ili se povremeno u nju vraćaju. Svaki će uzorak određenoga leksika, ma kako velik bio, »zalediti« zatečeno stanje. Taj se problem nastoji umanjiti povećanjem korpusa, pa se smatra da je leksičke analize neuputno raditi na korpusu manjem od milijun pojavnica.

Korpusi pisanoga jezika češći su i opsežniji od onih razgovornoga jezika, jer je transkripcija govornoga diskursa dugotrajniji posao, a metode označivanja arbitrarne, što može ograničiti lingvističku analizu.

Računalnim korpusom naziva se računalom potpomognut korpus, odnosno tekstualna baza podataka u strojno čitljivu obliku koja može biti obogaćena popratnim jezičnim informacijama (morfološkim, sintaktičkim, a katkad i fonološkim)[5], a konstruirana je tako da se može računalom pretraživati i organizirati po najrazličitijim parametrima. Računalni su korpusi najrasprostranjeniji izvor za kvantitativna jezična istraživanja, ali i alat lingvističkim istraživanjima drugih usmjerenja.

---

[2] Prvi sustavan referentni korpus jest BNC *(British National Corpus),* sastavljen 1992/93.

[3] Ako je potvrda samo jedna, entitet nije zanimljiv. Prava korpusna analiza kreće od 15 potvrda.

[4] Ne istražuju se paradigmatski, nego sintaktički odnosi, tj. u kojem se ko-tekstu, odnosno okolini taj jezični entitet realizirao.

[5] Što više dodataka pojavnice imaju u korpusu, to je korpus korisniji za različita lingvistička istraživanja.

Ciljevi, metodologija i alati korpusne lingvistike

Cilj je korpusne lingvistike steći dublji uvid u jezik i jezičnu uporabu istraživanjem korpusa pisanoga ili govornoga jezika. U načelu se ne istražuju izolirane rečenice, nego tekstovi različitih dimenzija.

Korpusna istraživanja, uza sam korpus, pretpostavljaju i suvremene alate i metodološke postupke koji jezičnu građu čine preglednom i dostupnom. Metodologijom korpusne lingvistike moguće je nadopunjivanje teorijskih tvrdnja i intuitivnih pretpostavka podatcima iz stvarne jezične uporabe u gotovo svim lingvističkim disciplinama i područjima.

Današnji uvid u korpus ne može se ni zamisliti bez pomoći računala, koja u lingvistici služe za prikupljanje i obradbu *istraživačkih podataka* (tj. *jezične građe* u obliku korpusa) te za provjeru istraživačkih hipoteza *računalnim modeliranjem* predmeta istraživanja (njegove strukture i odnosa u kojima se njegove sastavnice nalaze). Računalni modeli odgovaraju (strukturnim) modelima pojedinih (dijelova) *jezičnih podsustava* (u oblicima različitih modula za obradbu prirodnoga jezika: *taggeri, parseri*[6] itd.). Drugim riječima, računala u lingvistici služe za prikupljanje, uređivanje i pretraživanje jezične građe[7].

U okviru *jezičnih industrija* pojavljuje se potreba za *jezičnim alatima,* pomagalima kako jezikoslovcima u istraživanju jezika i informatičarima u dizajniranju i programiranju sustava za obradbu teksta, tako i svakodnevnim korisnicima računala pri manipulaciji tekstom. Naziv *jezična industrija,* nastao 1986., pokriva područja u kojima se s jedne strane razvija računalna pomoć za tradicionalna zanimanja primijenjene lingvistike (leksikografija, prevođenje, učenje stranih jezika), a s druge je strane usmjerena prema razvijanju novih aplikacija koje obrađuju prirodnojezične podatke (sustavi za prirodnojezična sučelja, analizu i sintezu govora, strojno indeksiranje, strojno prevođenje i sl.).

Programske potpore za obradbu korpusa[8] zaostaju za razvojem strojeva. Najbolje su razvijeni programi za izradbu konkordancija te za pretraživanje korpusa (npr. OCP – *Oxford Concordance Program*). Danas je za označivanje elemenata korpusa široko prihvaćen sustav XML *(Extensible Markup Language)*[9] i njegova inačica XCES *(Corpus Encoding Standard for XML),* a za zapisivanje pismèna (slova, znamenaka, interpunkcije) Unicode[10].

---

[6] Parser je računalni program za analizu rečenice do osnovnih sintaktičkih kategorija ili do leksikonskih unosaka.

[7] Tadić 1996.

[8] Lematizacija, razgraničivanje homografa ili morfološko i sintaktičko kodiranje.

[9] Nastao 1995/96. po uzoru na SGML *(Standard Generalized Markup Language)*.

[10] Unicode je globalni standard koji pokriva sva pismèna na svijetu (65 000), uključujući i kineska.

Jezične se tehnologije[11] bave različitim oblicima jezičnoga inženjerstva[12], obuhvaćaju primjenu jezičnoga znanja na interakciju između ljudi i strojeva, uvođenje automatizirane višejezičnosti u sustave i upravljanje informacijama koje su zabilježene u obliku prirodnoga jezika, a također uključuju[13]:

- prepoznavanje, razumijevanje i generiranje govora
- identifikaciju i provjeru govornika
- dizajn i analizu dijaloga
- prepoznavanje rukopisa
- crpljenje informacija i generiranje sažetaka
- strojno potpomognuto stvaranje i uređivanje teksta
- strojno potpomognuto prevođenje
- proizvodnju jezičnih resursa i alata za potporu.

Jezično se znanje pohranjuje u leksikonima, terminološkim bazama podataka, rječnicima vlastitih imena, gramatikama, semantičkim mrežama[14], itd. Jezični se resursi[15] sastoje ponajprije od korpusa (zbirke tekstova, računalni korpusi itd.)[16] ali i od rječnika (digitalni *on-line* i *off-line* rječnici, tezaurusi itd.)[17].

Jezičnotehnološki alati[18] za obradbu jezika na **fonemskoj razini** statistička su pomagala za proučavanje frekvencije fonema, odnosno grafema ili njihovih kombinacija. Za morfološku obradbu na **razini riječi** primjenjuju se označivači vrsta riječi *(POS /Part of Speech/ taggers),* morfosintaktički označivači *(MSD /Morpho-sintactic Description/ taggers)*[19] i lematizatori *(lemmatisers)*[20], a u leksikografiji alati za pretraživanje korpusa (konkordancije, kolokacijski upiti itd.)[21]. Na **semantičkoj razini** pri-

---

[11] Jezične tehnologije podrazumijevaju skup metoda i postupaka za preradbu prirodnoga jezika u sustave koji omogućuju korisnicima olakšanu uporabu (vlastitoga, prirodnoga) jezika u računalnome okružju (Tadić 2003).

[12] Jezičnim inženjerstvom smatra se primjena jezičnog znanja na razvoj računalnih sustava koji mogu prepoznavati, »razumjeti«, tumačiti i generirati ljudski jezik u svim njegovim oblicima (Tadić 2003).

[13] Tadić 2003.

[14] Najpoznatiju semantičku mrežu, *WordNet,* pokrenuo je 1990. u Princetonu G. Miller, a 1994. Europska je unija pokrenula *EuroWordNet*.

[15] Računalno pribavljene, pohranjene i podržane zbirke jezičnih podataka (Tadić 2003).

[16] v. http://www.hnk.ffzg.hr/jthj/korpusi.htm

[17] v. http://www.hnk.ffzg.hr/jthj/rjecnici.htm

[18] v. http://www.hnk.ffzg.hr/jthj/alati.htm

[19] Označuju rod, broj i padež za svaki imenički oblik.

[20] Svakoj pojavnici dodjeljuju njezinu lemu, polazni rječnički oblik.

[21] U tu skupinu pripadaju i alati za pisanje, pohranu, pretraživanje i distribuciju digitalnih rječnika (Tadić 2003).

mjenjuju se semantičke mreže[22] za značenje riječi i sustavi za prepoznavanje semantičkih uloga ili dubinskih padeža (agens, pacijens, instrument, itd.) u rečenici. Za obradbu jezika na **pragmatičnoj razini** služe se alati za istraživanja inteligentnih sustava, koji su potpomognuti i drugim medijima (slika, zvuk), a uključuju podsustave koji mogu uklopiti jezične rečenice u izvanjezični kontekst posredovan primjerice vizualno[23]. Iako je jedna od najranijih zamisli primjene računala bila strojno prevođenje *(machine translation, MT)*, još uvijek ne postoje sustavi za potpuno automatizirano visokokvalitetno **strojno prevođenje**[24]. Na području **strojno potpomognutoga učenja jezika** *(computer aided language learning, CALL)* postoje multimediji za individualno učenje, ali i sustavi za *on-line* poučavanje jezika.

Komercijalni proizvodi jezičnih tehnologija jesu provjernici *(chakers)*[25] pravopisa *(spelling-chakers)*, gramatike *(grammar-chakers)* i stila *(style-chakers)*, zatim rječnici (na CD-ima i *on-line*), sustavi za indeksiranje i sažimanje dokumenata, sustavi za crpljenje informacija i nazivlja, strojevi za diktiranje (sustavi *speech-to-text*) i sustavi za automatsko spikiranje (sustavi *text-to-speech*), korpusi te sustavi za strojno prevođenje.

Već je prije istaknuto kako je korpus temelj za svako istraživanje teksta, bez obzira na to promatra li se kao jezična građa ili nešto što se putem teksta tek ostvaruje. Korpusni se pristup, odnosno korpusna metodologija, lako može primijeniti u različitim lingvističkim disciplinama: fonologiji, morfologiji, sintaksi, sociolingvistici, kognitivnoj lingvistici. Sve su filološke znanosti po definiciji usmjerene na istraživanje tekstova, a računalnu obradbu tekstovne građe mogu primjenjivati i znanost o književnosti (kad konzultira činjenice teksta), povijest (kad zahtijeva uvid u dokumente, što se inače smješta u pomoćne povijesne znanosti poput arhivistike), dijelom i arheologija (kad proučava na/t/pise), ali i psihologija i sociologija. Računalna obradba teksta ne bi smjela pretendirati ni na kakvu epistemološku dimenziju. Njezina je svrha, kao istraživačkog alata, instrumenta ili pomagala, omogućiti znanstvenicima (i svima ostalima) ususstavljen i brz pristup velikim količinama teksta (Tadić 1996).

U Hrvatskoj je korpusna lingvistika najbolje rezultate ostvarila u okrilju kontrastivne analize te u nastavi stranoga jezika, osobito jezika struke (npr. frekvencijska istraživanja stručnih udžbenika[26]).

---

[22] U njima se značenja riječi opisuju njihovim dovođenjem u međusobne semantičke odnose, npr. *WordNet* (Tadić 2003).

[23] Tadić 2003.

[24] Strojno prevođenje jest prevođenje s jednoga jezika na drugi koje obavlja računalo, za razliku od strojno potpomognutoga prevođenja *(machine aided translation, MAT)* koje obavlja čovjek s pomoću računala. Najveći je korisnik strojnoga prevođenja EU. Primjenjuje se sustav za strojno prevođenje *Systran*.

[25] *Chaker* je sustav za provjeru točnosti i ispravnosti napisanoga teksta.

[26] M. Gačić i dr.

Korpusna lingvistika u svijetu

Od prvoga računalno potpomognutoga jezičnoga korpusa, milijunskoga korpusa engleskoga jezika, poznatijega kao Brownov korpus[27], suvremena korpusna lingvistika brzo je napredovala. U 1980-ima i 1990-ima brzi razvoj memorijskih mogućnosti osobnih računala rezultirao je izradbom velikih višemilijunskih korpusa, koji se često nazivaju korpusima druge generacije. Najpoznatiji je takav korpus stomilijunski korpus BNC *(British National Corpus),* koji je svojim sastavom reprezentativan za engleski jezik te služi kao uzor nacionalnim korpusima ostalih jezika. Korpusi treće generacije, odnosno korpusi budućnosti sadržavat će stotine milijuna riječi i bit će u komercijalnoj uporabi. To neće biti klasični korpusi, nego tekstualni arhivi (npr. *Oxford Text Archive*), koji bi mogli narasti i do veličine nacionalnih knjižnica.

John Sinclair ustvrdio je da su korpusi prve generacije dostatni samo za fonološke, morfološke i sintaktičke analize, da se različiti morfosintaktički obrasci mogu znatno bolje proučavati na većem broju primjera te da veliki korpusi iscrpnije pokazuju odstupanja. Tako se Sinclairov model Birminghamskoga korpusa[28] sastoji od osnovnoga zatvorenoga korpusa s oko 7 milijuna riječi, točno utvrđenih kriterija reprezentativnosti, te otvorenoga korpusa koji stalno raste i služi za provjeru rezultata dobivenih na osnovnom korpusu.

Većina je nacionalnih korpusa[29] reprezentativna za jezik u cjelini. Najbolje je sastavljen *Češki nacionalni korpus* (CNC) jer se jedini zasniva na sociološkim istraživanjima. Zanimljivo je da se *Američki nacionalni korpus* (ANC) počeo sastavljati tek 1999. te da se Korpus bosanskih tekstova sastavlja na Sveučilištu u Oslu. Na Matematičkom institutu u Beogradu, u sklopu projekta Matematička i računarska lingvistika, počeo se 1981. sastavljati *Korpus savremenoga srpskog jezika,* a 2002. postavljen je na Internet. Srpski *Projekat Rastko,* »elektronska biblioteka srpske kulture«, od 1997. arhivira tekstove iz srpske književnosti, ali prisvaja i neke hrvatske, kao i Kostićev korpus srpskoga jezika *(Corpus of Serbian Language by Đorđe Kostić)* koji ne daje objektivne rezultate o srpskom jeziku jer je uključeno i mnogo hrvatskih tekstova.

---

[27] Sastavili su ga 1967. N. Francis i H. Kučera na Sveučilištu Brown (Maryland, SAD).

[28] U sklopu projekta sveučilišta u Birminghamu i izdavačke kuće Collins Cobuild 1980-ih John Sinclair sastavio je *COBUILD (Collins Birmingham University International Language Database) Corpus* koji je poslužio kao glavni izvor za istoimeni jednojezični rječnik engleskoga jezika. To je prvi rječnik nekoga jezika napravljen isključivo na temelju građe iz korpusa i u njemu su značenja poredana po frekvenciji.

[29] *Poljski nacionalni korpus* (PNC), *Korpus slovenskoga jezika* (Fida), *Ruski nacionalni korpus* (RNC), *Slovački nacionalni korpus* (SNK), *Francuski nacionalni korpus (Trésor de la langue française informatisé)* i dr.

Korpusna lingvistika u Hrvatskoj

Kada je riječ o hrvatskim korpusima, ponajprije se misli na računalno potpomognute korpuse tekstova nastale na hrvatskome jeziku[30].

Prvi u nas računalno i iz lingvističkih pobuda obrađen korpus bio je korpus baroknoga epa *Osman* I. Gundulića, koji je, za boravka na Sveučilištu u Austinu, SAD[31], 1967. priredio, frekvencijski obradio i konkordancijama[32] popratio Željko Bujas. Pod vodstvom Rudolfa Filipovića, u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu, u okviru kontrastivnoga projekta pod nazivom *Yugoslav Serbo-Croatian – English Contrastive Project,* prvi je put u Hrvatskoj 1968. pokrenuta računalna obradba korpusa, za koju je bio zadužen Ž. Bujas. Brownov korpus prevođen je na tri standardne varijante hrvatskoga ili srpskoga, čime su se prvi put dobili i paralelni korpusi, a mogli su se pretraživati i engleski i prevedeni korpusi. To je ujedno bila i prva uporaba računala u svjetskoj kontrastivnoj lingvistici (Bujas 1975). Iste je godine u Zavodu za lingvistiku, pod vodstvom M. Moguša, pokrenut projekt *Jezik Marka Marulića,* koji je 1970. proširen i preimenovan u *Kompjutorsku analizu tekstova stare hrvatske književnosti* (Moguš 1975). Do 1981. konkordirana su hrvatska djela M. Marulića, djela B. Karnarutića, *Planine* P. Zoranića, *Ježupka* M. Pelegrinovića (3 inačice), djela H. Lucića i P. Hektorovića, *Hvarkinja* M. Benetovića, *Ranjinin zbornik,* komedije M. Držića, djela I. Bunića Vučića, djela P. Vitezovića Rittera, *Sveta Rožalija* A. Kanižlića, komedije T. Brezovačkoga, kajkavski ciklus pjesama F. Galovića, pjesme M. Pavleka Miškine i *Razvod istarski* te, neovisno o tom projektu, *Balade Petrice Kerempuha* M. Krleže. Tako dobiveni rezultati služili su ponajprije za nova kritička čitanja te za potvrđivanje autorstva gdje je ono bilo dvojbeno (npr. M. Marulić), ali i za istraživanja na svim jezičnim razinama, od leksičke do stilističke. U to je vrijeme hrvatska korpusna lingvistika potpuno pratila svjetska kretanja, tj. tzv. *Literary and linguistic computing.*

Također u Zavodu za lingvistiku od 1972. do 1975., pod vodstvom Ž. Bujasa, provođen je projekt *Englesko-hrvatski leksikografski korpus,* s polaznom nakanom obradbe teksta dvojezičnoga rječnika kao korpusa. To je dovelo do preokretanja i konkordiranja po hrvatskoj stožernici čitava Filipovićeva *Englesko-hrvatskoga rječnika*[33] (s više od milijun pojavnica), što je za to vrijeme također bilo jedinstveno računalnolingvističko dostignuće u dvojezičnoj leksikografiji.

---

[30] U Hrvatskoj se, naime, obrađuju ili su se obrađivali i korpusi i na drugim jezicima: npr. specijalističkim se engleskim korpusima bavio Boris Pritchard s Pomorskoga fakulteta Sveučilišta u Rijeci i Milica Gačić s Policijske akademije u Zagrebu.

[31] Bujas je ubrzo nakon toga na isti način priredio *Povratak Filipa Latinovicza* M. Krleže, a potom i *Suzanu* M. Marulića.

[32] Konkordancije su popisi riječi iz nekoga korpusa s ko-tekstnom okolinom (ono što slijedi i ono što prethodi) u kojoj su se riječi pojavile.

[33] Filipović, Rudolf: *Englesko-hrvatski rječnik*. Zora, Zagreb 1971.

Potreba za korpusom kojim bi se mogle proučavati pojave prisutne u jezičnoj sinkroniji, tj. korpusom reprezentativnim za suvremeni hrvatski jezik, potaknula je 1976. projekt *Korpus suvremenog hrvatskog književnog jezika*[34]. Osnovni je cilj bio sastavljanje jednomilijunskoga korpusa, nazvana Moguševim korpusom, obradba kojega se sastojala od abecednih i frekvencijskih rječnika pojavnica, konkordancija, potom i strojno potpomognute lematizacije[35] (Tadić 1992). Stjecajem različitih, a to znači i ratnih okolnosti, taj je korpus završen tek 1996., premda je njegova građa bila dostupna znatno prije. Bio je to prvi pokušaj u hrvatskoj lingvistici da se na temeljima reprezentativnoga korpusa počne ustavljivanje i istraživanje jezične građe. U vrijeme kad je zamišljen, bio je to prvi milijunski korpus nekoga slavenskoga jezika i bio je sukladan tadašnjim svjetskim kretanjima u sastavljanju korpusa.

Prvo naručeno korpusno istraživanje bio je frekvencijski korpus tekstova *Vjesnika* i *Večernjega lista* (opsega 130 000 pojavnica) koji je 1976. obradio Zorislav Šojat[36]. Na Odsjeku za informacijske znanosti Filozofskoga fakulteta Sveučilišta u Zagrebu, u prvoj je polovici 1980-ih sastavljen neuravnotežen korpus tekstova osnovnoškolskih udžbenika i novinskih tekstova (otprilike milijun pojavnica) pod vodstvom Damira Borasa, uz sudjelovanje Miroslava Kržaka i suradnika. Na tom je korpusu početkom 1990-ih istraživan probabilistički jezični označivač *(tagger)* za hrvatski.[37] Hrvatska je 1990–91. sudjelovala u međunarodnom višejezičnom leksikografskom projektu *(Multilingual lexicography project)* pod okriljem Vijeća Europe i vodstvom Johna Sinclaira. Voditeljica hrvatskoga segmenta bila je Maja Bratanić, a cilj projekta bio je proizvesti uzorak višejezičnoga rječnika kao modela novoga tipa instrumenta za strojno prevođenje[38]. Sudjelovanje je bilo moguće zahvaljujući ponajprije postojećemu Moguševu milijunskomu korpusu koji je dopuštao uvid u kontekst toliko potreban za pronalaženje prijevodnih ekvivalenata.

Zavod za lingvistiku uključio se 1995. u međunarodni projekt koji promiče međueuropsku suradnju na području jezičnih resursa – TELRI. Istodobno su se u Zavodu obrađivali korpusi J. Križanića (*Politika* 1988), I. Gundulića (ukupna djela 1989), M. Marulića (*Judita, Dijaloška djela, Suzana, Vartal,* u suradnji sa Splitskim književnim krugom, 1990–93), I. Mažuranića (*Smrt Smail-age Čengića* 1994) itd. Stručnjaci Zavoda za lingvistiku sudjelovali su i na projektu *Civilizacijska terminologija jugoistočne Europe,* u okviru kojega su Hrvatska akademija znanosti i umjetnosti i Austrijska akademija znanosti obradile korpus Katančićeva prijevoda *Svetoga pisma.*

---

[34] Autori projekta bili su Milan Moguš, Maja Bratanić, Vesna Muhvić-Dimanovski i Marko Tadić.

[35] Lematizacija je određivanje polaznog rječničkog oblika riječi.

[36] Šojat 1983.

[37] Žubrinić 1995.

[38] Bratanić 1992.

U sklopu projekta *Računalna obradba hrvatskoga jezika* Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu, pod vodstvom Marka Tadića, pokrenuo je 1996. sastavljanje *Hrvatskoga nacionalnoga korpusa* (HNK)[39] u dvjema sastavnicama: 30M (reprezentativni 30-milijunski korpus suvremenoga hrvatskoga jezika, s tekstovima nastalima od 1990) i HETA (Hrvatski elektronski tekstovni arhiv, s tekstovima starijima od 1990. i tekstovima koji ne odgovaraju zahtjevima reprezentativnosti, a sami čine znatne korpuse)[40]. U sastavljanju korpusa težilo se što boljoj reprezentativnosti jezične građe, što potvrđuje pet potkorpusa HNK-a kojima su obuhvaćene najvažnija izražajna područja pisanoga jezika: *novine, časopisi, knjige, beletristika, eseji* i *govori*. Od početka 2005., u okviru projekta *Razvitak hrvatskih jezičnih resursa,* omogućeno je naprednije pretraživanje HNK-a koje uključuje istodobnu pretragu s više riječi, tj. sa sintagmama, pretragu s pomoću dodatnih lingvističkih obavijesti (vrste riječi, gramatičke kategorije, automatsko pronalaženje kolokacija) i dr. Nova inačica HNK-a, iz proljeća 2008., podupire pretragu s pomoću lema i gramatičkih kategorija za dobivanje svih kombinacija *prijedlog + opća imenica*. HNK trenutačno obuhvaća 101,3 milijuna pojavnica.

U Institutu za hrvatski jezik i jezikoslovlje, pod vodstvom Dunje Brozović Rončević, 2005. pokrenut je projekt pod nazivom *Hrvatska jezična riznica*[41], u sklopu kojega se izrađuje nekoliko korpusa različitih razvojnih faza hrvatskoga jezika, digitaliziraju se objavljeni i rukopisni hrvatski rječnici, te se ujedno prikuplja i digitalizira građa za reprezentativni korpus hrvatskoga standardnoga jezika koji je podloga za izradbu Velikoga rječnika hrvatskoga jezika.

Korpusna leksikografija

Jezične analize, posebice semantičke, zahtijevaju sustavno selektiranu opsežnu jezičnu građu, pa su potrebu za višemilijunskim korpusima pokazala leksikološka istraživanja, koja su vrlo široko primijenjena prije svega u engleskoj leksikografiji. Osnovna je ideja takvih istraživanja promatrati pojavnice leksema u njihovu neposrednom okruženju i širem jezičnom kontekstu što omogućuju konkordancijski računalni alati[42]. Na temelju tako uređene jezične okoline moguć je pregledan uvid u gramatičke i leksičke značajke leksema s jasnim pokazateljima ukupnoga broja pojavnica pojedinoga leksema te frekvencije pojedinih značenja. Navedene značajke čine

[39] HNK u cijelosti je dostupan na internetskoj adresi http://www.hnk.ffzg.hr/

[40] Tadić 1997.

[41] Na internetskoj adresi http://riznica.ihjj.hr/index.html.hr moguće je pretraživati knjižni potkorpus, potkorpus tiskovina ili cjeloviti korpus.

[42] Najpoznatiji je oblik konkordancije tzv. KWIC (engl. *Keyword in Context*) koji je u korpusnim istraživanjima postulirao John Sinclair.

korpusnu jezičnu analizu primjerenom metodologijom za određivanje što točnijih frekvencijskih odnosa središnjega značenja i metaforičnih značenja pojedinoga leksema. Podatci o frekvenciji, odnosno o jezičnoj uporabi, do kojih se sustavno može doći s pomoću računalnoga korpusa, mogu biti dobro uporište za razmatranje utjecaja frekvencije pojedinoga leksema na opis strukture njegova značenja.

Mogućnosti iscrpnih semantičkih analiza povećavaju se pojavom velikih referentnih, odnosno za pojedini jezik reprezentativnih korpusa koji sadržavaju više stotina milijuna pojavnica. Postojanje tako opsežnih i uravnoteženih korpusa može umanjiti prigovore da su korpusne analize nedostatne. Računalni jezični korpus, zajedno sa svojom metodologijom, može poslužiti kao važan alat u semantičkoj analizi, a kako bi se dobio što potpuniji popis pojedinih značenja nekoga leksema. Pri traženju prototipnoga značenja nekoga leksema, možemo reći da za sve vrste leksičkih kategorija vrijedi tvrdnja da je prototip, zbog svojih kognitivnih prednosti, središnji među značenjima, i on predstavlja neposredno ili posredno izvorište ostalih značenja. Pri analizi frekvencijskih odnosa među pojedinim značenjima polisemnih leksema, uvid u neposredni i širi kontekst unutar referentnoga jezičnoga korpusa omogućuje grupiranje sličnih uporaba u pojedina značenja. Pri nejasnim razgraničenjima, tj. kada su značenja udaljena od prototipa, ovaj je postupak podložan subjektivnosti lingvista, što se često odražava u leksikografskim natuknicama polisema.

Mnogi su leksikografski pothvati (među kojima prednjače Sinclairovi) potvrdili prednosti korpusnoga pristupa u jednojezičnoj leksikografiji. Dok je pri sastavljanju jednojezičnih rječnika glavni izvor jednojezični korpus, u dvojezičnoj i višejezičnoj leksikografiji primjenjuju se **paralelni korpusi,** odnosno dvojezični ili višejezični korpusi koji sadržavaju niz tekstova na dvama jezicima ili na više njih, a koji su prijevodi s jednih na druge. Paralelnim se korpusima dvaju ili više jezika mogu smatrati korpusi načinjeni po istim načelima i odgovarajuće veličine[43]. Riječ je, dakle, o »usporedivim« (ne nužno »usporednim« korpusima). Paralelni korpusi omogućavaju leksikografima proučavanje kombinacije riječi i njihove prijevodne ekvivalente unutar konteksta u kojem se stvarno pojavljuju i pomažu im pri sastavljanju rječnika namijenjenih onima koji uče neki strani jezik. Na taj način korisniku mogu ponuditi važne podatke o nekim aspektima značenja riječi koji bi mu inače promaknuli te se usredotočiti na značenja i konstrukcije uporaba kojih je najrasprostranjenija. Digitalna priroda paralelnih korpusa omogućuje brže i lakše utvrđivanje prijevodnih ekvivalenata, sastavljanje elektroničkih rječnika[44] i jednostavno i brzo ažuriranje baze jezičnih podataka.

Paralelni korpusi također su bogat lingvistički resurs jer sadržavaju opsežnu količinu podataka o stvarnoj jezičnoj uporabi. Mnoga su područja njihove primjene:

[43] W. Teubert slikovito kaže da odgovarajući paralelni korpus mora biti »recipročan«.

[44] Paralelni se korpusi uspješno primjenjuju i u izgradnji dinamičkih rječnika *on-line.*

služe za razvoj sustava za strojno i strojno potpomognuto prevođenje, za kontrastivna i terminološka istraživanja, u glotodidaktici te u dvojezičnoj i višejezičnoj leksikografiji[45].

Rezultati obradbe nekoga korpusa mogu biti i **čestotnici** – glosari nekoga zatvorenog uzorka teksta ili tekstova s određenim brojem *pojavnica,* tj. svih riječi koje čine neki tekst (Bratanić 1992/93).

Prvi je čestotni rječnik u povijesti izrađen za njemački jezik. Izradio ga je ručno Johannes Käding, a izišao je u Steglitzu[46] 1897. pod nazivom *Häufigkeitswörterbuch der deutschen Sprache.* U njemu su se riječi navodile u obliku u kojem su pronađene u tekstu, a ne u kanonskome obliku. Početkom XX. st. počeli su se izrađivati i čestotnici drugih jezika: engleskoga[47], češkoga[48], francuskoga[49], mađarskoga[50], ruskoga[51] itd. Prvi čestotnik hrvatskoga jezika bio je uklopljen u rad I. Furlana *Raznolikost rječnika i struktura govora,* kao dodatak pod naslovom *Lista od blizu tri tisuće najčešćih riječi pisanog hrvatskog ili srpskog jezika,* ali je rad nažalost ostao u rukopisu.

Osim općih čestotnika, koji trebaju biti reprezentativni za jezik u cjelini, postoje i čestotnici različitih »podjezika« (društvenih skupina, struka, određenih razdoblja, geografskih područja itd.), idiolekata pojedinaca (čestotnici sabranih djela pojedinih književnika, političara, novinara, itd.) te čestotnici pojedinačnih tekstova (npr. *Biblije*).

Standardni oblik čestotnika donosi popis *različnica* (svih različitih riječi)[52] nekoga korpusa po silaznome frekvencijskom slijedu, dakle od najučestalije riječi do one ili onih koje se u korpusu pojavljuju samo jednom. Uz njih se najčešće bilježi njihov »rang« na čestotnoj listi te njihova apsolutna i relativna frekvencija. Ti se podatci mogu predočiti zasebno i za pojedine potkorpuse (s obzirom na žanrovsku, stilsku, tematsku ili neku drugu odrednicu) glavnoga korpusa. Uz popis natuknica prema čestoti, čestotnici obično donose i abecedni rječnik, u kojem se natuknice s pridruženim podacima o čestoti navode abecednim redom.

---

[45] Simeon 2002.

[46] Reprint-izdanje izišlo je 1963. u Hamburgu.

[47] Prvi poznati čestotnik engleskoga jezika, *A Measuring Scale for Ability in Spelling* L. P. Ayresa, izišao je u New Yorku 1915.

[48] Najstariji češki čestotnik, *Základní studie k českému tesnopisu* J. Sedláčka, ostao je u rukopisu.

[49] Najstariji je francuski čestotnik iz 1920–21. i to je rječnik 1. broja novina *Le Temps.*

[50] Prvi mađarski čestotnik, *Neue ungarische Wörterstatistik,* izradio je 1942. za potrebe stenografije Z. Nemes.

[51] Prvi čestotnik ruskoga jezika, *Rusko-český slovník najdůlžitejších slov pro četbu sovetského tisku,* izradili su studenti ruskoga jezika u Pragu 1951. pod vodstvom F. Malira.

[52] U lematiziranom se čestotniku broj natuknica smanjuje jer se svi različiti oblici istoga leksema svode na kanonski oblik. Takvim se postupkom dobiva dorađeniji oblik rječnika, ali se mogu izgubiti dragocjeni podatci o učestalosti različitih oblika pojedine riječi (Bratanić 1992/93).

Tako se *Hrvatski čestotni rječnik*[53], koji je rezultat obradbe Moguševa milijunskoga korpusa, sastoji od rječnika lema cijeloga korpusa navedenih po čestotnom redu, čestotnika pojedinih potkorpusa[54] i abecednoga rječnika lema, koji uz osnovnu natuknicu donosi podatke o njezinoj apsolutnoj i relativnoj frekvenciji, te popis svih njezinih oblika s frekvencijskim pokazateljima.

Zaključak

Korpusi će biti sve veći i sve mnogobrojniji, a osobit se napredak očekuje u sastavljanju korpusa govornoga jezika, što podrazumijeva i razvoj instrumenata i alata za njihovu analizu. Mogućnosti primjene korpusne lingvistike u suvremenoj leksikografiji, te osobito u području strojnoga prevođenja, iznimno je golem. Zašto to ne iskoristiti?

LITERATURA

Bratanić, Maja: Korpusna lingvistika ili sretan susret. *Radovi Zavoda za slavensku filologiju* 27(1992).

Bratanić, Maja: Mjesto čestotnika u jezičnom opisu. *Filologija* 1992/93, knj. 20–21.

Bujas, Željko: Computers in the Yugoslav Serbo-Croatian – English Contrastive Project. *Bilten Instituta za lingvistiku* 1(1975).

Kennedy, Graeme: *An Inroduction to Corpus Linguistics*. Longman, London–New York 1998.

Lončarić, Mijo: O čestotnim rječnicima i čestotniku hrvatskoga književnog jezika. Suvremena lingvistika 15(1977).

Moguš, Milan: Kako su se Marulićeva djela našla u kompjuteru. *Bilten Instituta za lingvistiku* 1(1975).

Simeon, Ivana: Paralelni korpusi i višejezični rječnici. *Filologija* 2002, knj. 38–39.

Šojat, Zorislav: *Čestotni rječnik Vjesnika i Večernjeg lista*. Zagreb 1976.

Tadić, Marko: Od korpusa do čestotnog rječnika hrvatskoga književnog jezika. *Radovi Zavoda za slavensku filologiju* 27(1992).

Tadić, Marko: Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika* 22(1996)41–42.

Tadić, Marko: Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. *Suvremena lingvistika* 23(1997)43–44.

Tadić, Marko: Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika. *Filologija* 1998, knj. 30–31.

Tadić, Marko: *Jezične tehnologije i hrvatski jezik*. Ex libris, Zagreb 2003**.**

Žubrinić, Tomislava: *Mogućnosti strojnog označavanja i lematiziranja korpusa tekstova hrvatskoga jezika* (magistarski rad), Filozofski fakultet Sveučilišta u Zagrebu, Zagreb 1995.

[53] Moguš, Milan, Bratanić, Maja, Tadić, Marko: Hrvatski čestotni rječnik. Zavod za lingvistiku, Školska knjiga, Zagreb 1999.

[54] Glavni se korpus sastoji od pet potkorpusa sačinjenih od uzoraka proznih tekstova, poezije, dramskih tekstova, novina i srednjoškolskih udžbenika.

CONTRIBUTION OF CORPUS LINGUISTICS TO MODERN LEXICOGRAPHY

Iva Klobučar Srbić
The Miroslav Krleža Lexicographic Institute, Zagreb

SUMMARY: Corpus linguistics has taken an irreplaceable position in modern lexicographic practice. This paper deals with the corpora, in particular computer corpora, language technologies and tools, and the application of corpus linguistics, especially in modern lexicography. A special chapter is dedicated to corpus linguistics in Croatia, its inception, as well as contemporary projects and achievements.

Keywords: *corpus, corpus linguistics, language technologies, corpus lexicography*