Prethodno priopćenje

https://www.doi.org/10.17234/9789533791555.19

Ivana Filipović Petrović

Hrvatska akademija znanosti i umjetnosti, Zagreb
Zavod za lingvistička istraživanja
ifilipovic@hazu.hr

Što je hrvatskoj leksikografiji korpus[1]

Suvremena teorijska leksikografija, kao važno obilježje rječnika, ističe pouzdanost jezičnih podataka na kojima se rječnik temelji (Atkins i Rundell 2008; Hanks 2012). Da bi bili pouzdani, smatra se da ti podaci trebaju vrlo približno odgovarati onome kako govornici koriste i razumiju jezik u stvarnoj komunikaciji. Nekoliko je izvora podataka na kojima se mogu temeljiti rječnici, ovisno o njihovoj namjeni i tipu: introspekcija, odnosno oslanjanje na vlastiti mentalni leksikon, jezičnu intuiciju i kompetenciju, zatim rječnici *prethodnici*, pa ručno prikupljena građa putem anketa i terenskih ispitivanja ili iz književnih djela i naposljetku, u digitalno doba, višemilijunski računalno pretraživi jezični korpusi. U ovome radu bavimo se leksičkim izvorima na kojima se temelje jednojezični rječnici u hrvatskoj suvremenoj leksikografiji. Konkretno, proučava se stav hrvatskih leksikografa prema *sirovoj* građi za rječnik koji se u njima ogleda. Pritom se ukazuje na nemalu ulogu introspekcije kao izvora podataka u hrvatskim rječnicima te stanovite zadrške prema dostupnim računalnim korpusima za hrvatski jezik. Naposljetku, na tragu težnje za pouzdanim rječnikom koji odražava stvarnu upotrebu, oblikuju se uvjeti koje treba ispuniti takav rječnik, a koji hrvatska leksikografija tek treba proizvesti.

Ključne riječi: suvremena leksikografija, računalni korpusi, opći jednojezični rječnici, izvori

1. Uvod

Rječnici se često povezuju s pojmom autoriteta: velike su povijesne rječnike sastavljali jezični i uopće društveni autoriteti, a temeljili su se na citatima iz djela književnih autoriteta. U doba računalnih korpusa pojavio se još jedan autoritet – autoritet jezičnih dokaza. Taj je autoritet mjesto našao u deskriptivnoj leksikografiji čiji

[1] Naslov rada inspiriran je naslovom tematski bliskoga rada Što je hrvatskoj leksikografiji natuknica Branke Tafre (2012). Zahvaljujem prof. Tafri na suglasnosti za takav odabir naslova.

se začeci bilježe još 1728. kada je Ephraim Chambers zapisao da leksikograf, baš kao i primjerice povjesničar, dolazi *poslije*, što znači da daje opis nečega što se već zbilo (Mugglestone 2016: 555). Jezične dokaze o tome *što se zbilo* pružaju korpusi, koji se danas smatraju najvažnijim empirijskim temeljima u leksikografiji i uopće lingvistici (Kupietz 2016). Suvremena deskriptivna leksikografija želi u korpusu saznati kako se jezik tipično, uobičajeno koristi. Premda se shvaćanje da dobri pisci odražavaju dobru upotrebu jezika i da je to ono što rječnik treba oslikati zadržalo prilično dugo, iz modernih korpusa sve više izostaju pjesnička djela, dok se prozna smanjuju. Dakako, vrijednost književnih citata nije upitna, no jezik velikih književnika teško da je reprezentativan za svakodnevnu običnu upotrebu, pa se postavilo pitanje njihove korisnosti – i za leksikografa dok pokušava izvući značenje ili značenja riječi, ali i za korisnika rječnika. Od 2000. raste popularnost korpusa koji se baziraju isključivo na tekstovima s interneta, pa iz inicijative *Web as Corpus Kool Ynitiative* (WaCky) (Baroni i sur. 2009) nastaju korpusi za engleski (ukWaC), njemački (deWaC), talijanski (itWaC) i druge jezike te između ostalih i – hrvatski (hrWaC). Kako je Sinclair istaknuo 1987. na početku projekta Coublid, korpus je izum pomoću kojega onaj tko ga koristi promatra živi jezik.

U ovome radu, istražuje se stav hrvatskih leksikografa – autora hrvatskih suvremenih općih rječnika – prema korpusima, odnosno prema izvorima jezične građe za rječnike koji se ogleda u njihovim predgovorima i leksikografskoj obradi. Uočava se, naime, da je hrvatska leksikografija kontinuirano sklona autoritetu pojedinca i eventualno autoritetu jezika književnosti, dok autoritet stvarne upotrebe jezika koja se ogleda npr. u hrvatskom mrežnom korpusu hrWaC, nije dovoljno snažan da bi služio kao kriterij za uvrštavanje u rječnik. Takav je pristup povezan sa shvaćanjem rječnika kao mjesta na kojem će se poštivati standard i dobra upotreba u skladu s propisanom normom. Ali ne samo to, s time je u vezi i vjerovanje da za svaku riječ postoji njezino zadano pravo značenje koje čuva i daje – rječnik (usp. Barnhart 1962; Béjoint 2010). U tom se smislu ponekad u jezičnim savjetima rječnik koristi kao sredstvo koje legitimira određene oblike i značenja ako ih bilježi, što Starčević (2016: 94–95) naziva ideologijom rječničke potvrde, u kontekstu drugih preskriptivističkih ideologija (više i u Starčević i dr. 2019). Istraživanje u ovome radu pokazalo je da normativni aspekt svakako ima svoj pečat u hrvatskim općim rječnicima, što utječe na nekorištenje ili djelomično korištenje korpusa u izradi rječnika, koje se opisuje kao nereprezentativne. Pitanje reprezentativnosti nadalje otvara raspravu o tome je li uopće moguće obuhvatiti jedan jezik u cjelini u jednome korpusu, je li to uopće namjena korpusa, i kada bi se to postiglo, bi li takav korpus zadovoljio jezikoslovce u ulozi leksikografa dovoljno da svoju introspekciju ostave po strani.

Rad započinjemo osvrtom na tradicionalnu leksikografiju i shvaćanje uloge rječnika tijekom povijesti, a zatim prelazimo na suvremene lingvističke struje i pojavu korpusnoga pristupa u leksikografiji. U središtu je rada analiza suvremenih hrvatskih općih

rječnika, odnosno njihova stava prema jezičnoj građi na kojoj se temelje, a na kraju slijedi zaključak.

2. Introspekcija autoriteta

Povijest leksikografije, svjedoče o tome liste riječi s kratkim objašnjenjima koje postoje u svim drevnim kulturama, traje od prvih pisanih tragova. Potreba za sastavljanjem neke vrste rječnika oduvijek je postojala, ali ne samo to: privlačila je mnoge genijalne umove iz raznih drugih struka da se upuste u leksikografiju (v. Filipović Petrović 2018: 27–28). Jezični podaci kao izvori za rječnik svim su velikim leksikografima bili važni: osim introspekcije pojedinca koji sastavlja rječnik i koji je obično imao veliko znanje najčešće i o jeziku i o svijetu, najvažnijim izvorom bila su književna djela. Najznačajniji engleski leksikograf, Samuel Johnson, svoj je rječnik, preteču znamenitog *Oxford English Dictionaryja,* temeljio na velikim djelima engleske književnosti i većinu je značenja u rječniku potkrijepio citatima iz tih djela. I hrvatski su rječnici, poput Della Bellina2 i Stulijeva,3 slijedili primjer rječnika *velikih* europskih jezika i uključili su književne citate, odnosno primjere upotrebe riječi u djelima poznatih pisaca u svoje rječnike. Ta je svojevrsna korpusna metoda izrade rječnika, pri čemu se leksikografska rješenja temelje na primjerima iz upotrebe u književnim djelima, kulminirala u Benešićevu *Rječniku hrvatskoga književnoga jezika od preporoda do I. G. Kovačića* čiji izvori broje 466 djela hrvatske književnosti. Dakako, taj pristup nije isključio introspekciju iz leksikografije, dapače, prikupljanje građe obilježeno je intuicijom i jezičnim stavovima svojih ekscerptora.4 Štoviše, Akademijinu je rječniku upravo koncepcija, odnosno izbor građe koju će uključivati,5 donijela brojne prigovore, kasnije preinake i zapravo je ostao trajno obilježen jezičnom politikom koja ga je prožela.

Prvotna je namjera uključivanja književnih citata u rječnike bila da se pokaže postojanje riječi i da su je koristili veliki književnici. Međutim, kada je u razdoblju prosvjeti-

[2] U trojezičnome rječniku *Dizionario italiano, latino, illirico* (1728) Della Bella je zabilježio popis od dvadesetak autora iz čijih je djela ekscerpirao rječničku građu i kod gotovo svake natuknice književnu potvrdu.

[3] U Stullijevu trodijelnome rječniku nalazi se i popis literature od sto pedeset djela.

[4] Benešićev rječnik, premda u ideji bez normativnih pretenzija, također nije izmaknuo izvanjezičnim motivima pa je njegov autor, i sam štokavac, pri izradi nomenklature za rječnik u znatno većoj mjeri birao štokavske izraze, iako u 466 odabranih djela ima i kajkavskih pisaca (usp. Schubert 2018).

[5] Prvim je urednikom toga rječnika i autorom koncepcije bio Đuro Daničić koji je, djelujući u ozračju jugoslavenske ideologije i zamišljenoga hrvatsko-srpskog jezičnog jedinstva koje je forsiralo novoštokavski purizam, zanemario kajkavsko narječje, kajkavski književni jezik i hrvatske pisce 19. stoljeća, zbog čega su rječniku upućivani brojni prigovori. Kasniji su urednici uvodili nove pisce i više pazili na uočene nedostatke.

teljstva na scenu stupila potreba za propisivanjem jezičnog standarda, uz rječnike se počeo, ponekad i doslovno,[6] vezati pojam autoriteta. Zamišljeno je da baš pisci i književna djela iz kojih su crpljeni citati za rječnik predstavljaju te autoritete pomoću kojih će se odrediti pravila za jezike i učiniti ih čistima od svega nepotrebnog. Bio je to začetak ideje o rječniku kao autoritetu koji sugerira učenost i preciznost koja je kasnije, u doba formiranja nacionalnih jezika, postala neupitno obilježje rječnika koje ga prati i danas.[7] Kada je riječ o jezičnim izvorima za rječnik, možda najveća šteta učinjena je sredinom 18. stoljeća kada se uključivanje citatnih potvrda okarakteriziralo kao nepotrebno zauzimanje prostora[8] u rječniku čija je uloga regulirati jezik. Ipak, književni su citati sve do 1980-ih bili glavni oblik empirijskih jezičnih podataka dostupnih leksikografima koji su željeli takve jezične dokaze. Štoviše, nacionalne leksikografije većine europskih jezika upustile su se u izradu velikih općih rječnika temeljenih na još većim kartotekama s građom iz književnosti. Doduše, ručno prikupljanje takvih kartoteka redovito bi nadživjelo barem dvije generacije leksikografa i samo u najboljem slučaju takvi su rječnici bili završeni, često desecima godina nakon početka rada. Iscrpljeni takvim mukotrpnim radom, ali i zahvaljujući računalnoj revoluciji koja je bila na pomolu, leksikografi i leksikografija našli su se na prekretnici: ručno skupljanje citata počet će zamjenjivati računalni korpusi. Međutim, djelomično i zbog spomenutih snažnih stavova o normativnoj ulozi rječnika u društvu, korpusni pristup u leksikografiji nije odmah, a katkad ni uopće, postao norma u pojedinim nacionalnim leksikografijama, uključujući i hrvatsku. O korpusnome pristupu i promjenama koje je donio u leksikografiju, govori se u nastavku.

3. Zaokret u lingvistici i posljedice za leksikografiju

Na prvi se pogled može činiti da su dvije discipline, lingvistika i leksikografija, dio svoje povijesti provele koračajući međusobno odvojenim putevima (usp. Bratanić 1991) bez posebnog razloga. Međutim, ako razmislimo o riječima Bernarda Quemade (1972: 427) da je svaki leksikografski rad odraz određene lingvističke teorije koju leksikograf više ili manje svjesno primjenjuje, shvaćamo da je ta odvojenost zapravo znakovita. Smatrajući leksikon najmanje važnim dijelom jezika, posebno u odnosu na sintaksu i fonologiju, strukturalistički se lingvisti nisu mnogo bavili leksikografijom, a iz istih razloga na rječnike nije imala utjecaja ni generativna lingvistika (usp. Hanks

[6] Npr. španjolski rječnik *Diccionario de autoridades* (1726).

[7] O rječniku kao autoritetu vidi npr. Landau (2001).

[8] Prigovori koje su recenzenti rukopisa *Rječnika hrvatskoga književnog jezika od preporoda do I. G. Kovačića* uputili njegovu autoru Juliju Benešiću odnosili su se između ostalog upravo na preveliku količinu citata koji zauzimaju previše mjesta (Benešić 1985). To pokazuje da je takav stav bio raširen, premda ga (nasreću) nije dijelio i Julije Benešić koji nije odustao od svojih brojnih citata.

1990: 31; Béjoint 2010: 264–265). Do promjena je došlo zaokretom prema leksikonu i semantici u smjeru proučavanja semantičkih obilježja pojedinih riječi, njihove uloge u diskursu i njihovih veza s drugim riječima (npr. Cruse 1986). Također, kognitivna je lingvistika u središte interesa dovela uporabni model jezika (Langacker 1988, 2000; Kemmer i Barlow 2000: xv) prema kojemu je za shvaćanje prirode i funkcioniranja jezika najvažnije u obzir uzeti njegovu upotrebu. To je otvorilo vrata razvoju empirijskoga pristupa jezičnim podacima, odnosno ispitivanju i zaključivanju na temelju proučavanja stvarne jezične upotrebe – korpusnoj lingvistici.

Paralelno s time, stavovi leksikografa o tome što je rječnik, odnosno što bi trebao biti, i što treba raditi leksikograf kretali su se u smjeru deskriptivnosti utemeljenoj na jezičnim podacima. Digitalno doba i pojava velikih računalnih korpusa to su učinili mogućim. Za autore priručnika *Oxford Guide for Practical Lexicography*, engleske leksikografe Sue Atkins i Michaela Rundella, rječnik je opis vokabulara koji koriste govornici nekog jezika (2008: 3). Drugim riječima, taj opis sasvim približno odgovara onome kako ljudi koriste i razumiju jezik u stvarnoj komunikaciji, bilo da pišu roman ili poslovni dopis, čitaju novine ili razgovaraju međusobno. Da bi se to ostvarilo, vrlo su važni izvori jezičnih podataka za rječnik na temelju kojih se ostvaruje njegova pouzdanost. Naime, leksikografi žele biti sigurni, kad sastavljaju rječnik, da znaju kako govornici koriste jezik i riječi i da će njihov rječnik korisnicima biti pouzdan. Stoga se dosta rasprava posvetilo polazišnoj točki za ostvarenje te pouzdanosti: izvoru jezičnih dokaza na kojima se temelje zaključci o jeziku i prenose u rječnik.

Jedan je od izvora introspekcija koju smo spomenuli: proces u kojem značenje pojedine riječi tražimo u našem mentalnom leksikonu, dakle oslanjanje na vlastitu jezičnu intuiciju i kompetenciju. Kritika toga pristupa (Schütze 1996, 2016) upozorava da, premda ga stalno koristimo, takav postupak nije dovoljan za pouzdan rječnik jer je naše jezično znanje individualno i nužno nepotpuno. Autor iz psiholingvističke perspektive propituje što se događa kada iz naše intuicije pokušavamo izmamiti jezični podatak i što nam zapravo taj podatak govori u lingvističkom smislu. Raspravljajući o intuiciji izvornih govornika kao izvoru jezičnih podataka, sažima tri problema: najprije, ti podaci nisu dobiveni sustavno, organiziranim bilježenjem, nego sporadično i po potrebi, zatim, podaci se često prekrajaju (prihvaćaju ili odbacuju) kako lingvistu odgovara i konačno, podrazumijeva se da je intuicija jednog izvornog govornika pouzdana bez namjere da se podaci sistematično prikupljaju (2016: 52–53). Tako nastaju, zaključuje Schütze, rječnici intuicije, koji nemaju mnogo veze sa svakodnevnom upotrebom i razumijevanjem jezika kod govornika. Naime, kada su se pojavili prvi veliki računalni korpusi za engleski jezik i kada su omogućena empirijska istraživanja, spoznalo se da intuicija izvornog govornika o učestalosti pojedinih riječi, kolokacija i značenja uopće nije pouzdana (Hanks 2012). Nadalje, pitanje pouzdanosti rječnika javljalo se i u raspravama o pozicioniranju leksikografije kao znanstvene discipline. John Sinclair (1984) smatrao je da je moguće leksikografiju etablirati kao znanstvenu disciplinu, ali

da se za to moraju ispuniti neki uvjeti. Između ostalog vrlo problematičnom smatra raširenu praksu ciljanog smišljanja primjera,[9] nazivajući to vrstom varljive igre koju leksikografi igraju. Tvrdi se kako smišljeni primjeri ilustriraju upotrebu riječi, no potrebno je osvijestiti da upotreba nije valjana ako je smišljena, već samo ako je stvarna (Sinclair 1984: 3–4).

Ostali izvori jezičnih podataka za rječnik poput anketa i eksperimenata korišteni su za mnoge rječnike, no kritika je isticala da poteškoće donosi činjenica da je ispitanik svjestan namjere procjene nekog jezičnog fenomena i zbog toga se stvara sumnja u to hoće li se dobiti stvarna upotreba, odnosno važno je uložiti napor u skrivanje namjera pitanja (v. npr. Labov 1972; Atkins i Rundell 2008). Naposljetku, ručno skupljanje citata iz književnosti kao dugotrajnu i zastarjelu metodu prožetu introspekcijom istisnula je pojava računalnih višemilijunskih korpusa.

Suvremena leksikografija je, dakle, zauzela stav da nijedan rječnik utemeljen na suvremenim principima ne treba propisivati upotrebu, nego informirati na temelju jezičnih podataka o upotrebi. Ideja da rječnici postoje da bi podržavali određeni jezični standard i presuđivali koja je dobra, a koja loša upotreba određene riječi fundamentalno je suprotna ideji korpusne lingvistike. Smisao upotrebe korpusa je upravo u tome da se izbjegne prosuđivanje tekstova i biranje prema tome odobrava li ih neki pojedinac ili ne. Rosamund Moon ističe da je, zahvaljujući podacima iz korpusa zadatak leksikografije, u posljednjih tridesetak godina iz opisa idealnoga jezika prerastao u idealizirani opis jezika u upotrebi (2008: 315). Nadalje, zahvaljujući tehnologijama digitalnog doba, područje elektroničke leksikografije preplavilo je leksikografske projekte, rječnike, a i korisnike. Sučelja e-rječnika dizajniraju se na način da funkcionalno prikazuju rječnički sadržaj, na tragu funkcionalne teorije leksikografije danske Aarhuške škole[10] (Leroyer 2007: 110; Nielsen i Mourier 2007). Funkcionalna leksikografija podrazumijeva da je proizvodnja rječnika vođena korisničkim potrebama s ciljem pružanja znanja potrebnog za komunikaciju u određenim situacijama u kojima se korisnik može naći (Fuertes-Olivera 2009). Posebna se pažnja usmjerila na mogućnosti pružanja različitih dodatnih leksičkih informacija, kao što su sinonimi, širi kontekst, informacije o čestoj i specifičnoj upotrebi. Pokazat će se u nastavku, računalni su korpusi neizostavni dio te priče.

3.1. Korpusni pristup u suvremenoj leksikografiji

Kao posljedica spomenute težnje za pouzdanošću jezičnih podataka, ali i potrebe za velikom količinom tih podataka kako bi se mogao steći uvid u čestotu upotrebe razli-

[9] Javlja se i u nekim hrvatskim rječnicima, npr. *Hrvatskom frazeološkom rječniku* (2014: 15).

[10] Aarhuška je škola osnovana 1996. na Sveučilištu u Aarhusu u Danskoj, u sklopu kojeg postoji Centar za leksikografiju na čelu s Henningom Bergenholtzom. Zastupaju funkcionalnu teoriju leksikografije smatrajući je dijelom informacijskih znanosti, fokusirajući se na različite funkcije rječnika (Bergenholtz, Nielsen i Tarp 2009).

čitih značenja i oblika riječi i višerječnica, postupno su izrađivani nacionalni korpusi pojedinih jezika koji su postali nezaobilazni u njihovoj leksikografskoj praksi (usp. Klobučar Srbić 2008). Korpusna je lingvistika razvila svoju teoriju i metode (Biber i sur. 1998; Tognini-Bonelli 2001; Meyer 2002; McEnery i Gabrielatos 2006) kojima ćemo ovdje posvetiti nekoliko dodatnih redaka zbog njihove važnosti za temu rada.

Uz objašnjenje korpusnoga pristupa često se kaže i da je suprotan tumačenjima jezika zasnovanima na jezičnoj intuiciji pojedinca, a pritom se misli na to da intuicija ne služi za to da se iz nje crpe podaci za analizu i ne dominira nad podacima dobivenim empirijskim istraživanjem. Također, zbog kritika upućenih eventualnoj neobuhvatnosti jezika u korpusima, treba istaknuti da korpusna lingvistika ispituje uzorke jezične upotrebe, doduše ogromne, ali uzorke, kao dijelove statističkog skupa koji predstavljaju njegovu cjelinu, jer potpunu ukupnost jezika u jednom korpusu nije moguće ispitivati. Odnosno, određene se ukupnosti mogu ostvariti u specijaliziranim korpusima, ako se primjerice izradi korpus cjelokupnog opusa nekog autora ili periodike određenog razdoblja. Temeljno je mjerilo istraživanja u korpusnome pristupu frekvencija, dakle podaci o čestoći pojavljivanja i supojavljivanja leksičkih jedinica.

Upotreba korpusa u literaturi uglavnom se dijeli na dvije, odnosno tri metode: *corpus-driven* i *corpus-based* (Tognini-Bonelli 2001), odnosno *corpus-illustrated* pristup (Tummers, Haylen i Geeraerts 2005; Geeraerts 2006: 36–37). Za hrvatski se jezik u *Hrvatskom terminološkom priručniku* (2009: 92) navode termini: pristup utemeljen na korpusu i pristup vođen korpusom. Tognini-Bonelli (2001) pod *corpus-based* pristupom smatra onaj koji se korpusom služi kako bi se provjerile unaprijed postavljene hipoteze i razlikuje se od pristupa koji hipoteze postavlja isključivo na temelju rezultata korpusne analize. Metoda kojom se rječnik temelji na korpusu uključuje korištenje korpusa za tumačenje, provjeravanje i pronalaženje primjera za teorije i opise koji su oblikovani prije nego što su se veliki računalni korpusi pojavili u jezikoslovlju (Tonini-Bonelli 2001: 65). Korpusom vođena metoda suvremeni je izazov i za lingvistiku i za leksikografiju: svaki je korak u tome smjeru korak prema potencijalnom postavljanju novih hipoteza i nepotvrđivanju starih. Naime, ta promjena pristupa, možemo reći i stava prema jeziku, koju su indirektno donijeli računalni korpusi, za lingviste je neumoljiva: jezične dokaze koje dobivamo iz korpusa možemo prihvatiti ili argumentirano odbiti, ali ih ne možemo – ignorirati (v. Tonini-Bonelli 2001: 84–100). Način upotrebe korpusa koji Tummers, Haylen i Geeraerts (2005) nazivaju *corpus-illustrated* podrazumijeva korpus kao repozitorij primjera. Drugim riječima, primjeri koji se koriste u jezičnoj analizi više nisu smišljeni iz glave lingvista, nego se posežе za korpusom, ali s namjerom da primjeri oslikaju određenu jezičnu pojavu koja i dalje proizlazi iz vlastitog jezičnog znanja pojedinca, dakle introspekcija je i dalje dominantan pristup jezičnog građi (Geeraerts 2006: 36–37). Premda se kao slabosti *corpus-based* i *corpus-illustrated* pristupa navodi njihova utemeljenost u intuiciji, neki autori smatraju da je problematičnije ono što slijedi nakon pretrage rezultata. Primjerice, kada dođe do nepo-

dudaranja između očekivanih teorijskih postavki i podataka koje korpus pokazuje, neki od postupaka su izoliranje podataka koje korpus daje, a ne odgovaraju teoriji, dakle prešućivanje, ili njihovo prilagođavanje, što podrazumijeva ulaženje u korpus i njegovo preinačavanje (Tonini-Bonelli 2001: 68–71; McEnery i Gabrielatos 2006: 35–36).

Još u začecima korpusne revolucije hrvatska je leksikografkinja i jezikoslovka Maja Bratanić istaknula da će se korpusnim pristupom, dakle promatranjem tipične i česte jezične upotrebe i njezinim minucioznim opisom stvoriti iscrpni jezični priručnici, dok će put koji polazi od intuicije ostati rezerviran za interpretaciju jezične kreativnosti, iščitavanje nenapisanog i dohvaćanje impliciranog (1997: 11). Premda vrlo smislene i optimistične, čini se da se te misli nisu ostvarile, o čemu raspravljamo u nastavku.

4. Hrvatski rječnici i suvremena leksikografija

4.1. Predkorpusni rječnici

Premda je hrvatska leksikografija od početaka pratila korak s leksikografijama drugih europskih jezika i prošla je sve razvojne faze kao i većina drugih – rječnici su kroz povijest služili za podučavanje i tumačenje, za jezični purizam i jezičnu politiku, za pokazivanje da su određenu riječ koristili slavni književnici, za učenje stranih jezika, za usustavljivanje drugih disciplina, poput nazivlja, kao i drugih znanosti, poput prava ili pomorstva – primijećeno je da suvremena rječnička produkcija u nas ne prati dovoljno mogućnosti digitalnog doba i dostignuća elektroničke leksikografije (Klobučar Srbić 2008; Štrkalj Despot i Möhrs 2015; Parizoska i Filipović Petrović 2016; Filipović Petrović i Parizoska 2017; Filipović Petrović 2018).

Tri se rječnika danas u Hrvatskoj rade na temelju ispisane kartoteke: Benešićev *Rječnik* koji obuhvaća književni jezik od polovice 19. do polovice 20. stoljeća, zatim *Rječnik hrvatskoga kajkavskoga književnog jezika* koji obuhvaća stariju i noviju kajkavsku književnost te *Rječnik crkvenoslavenskoga jezika hrvatske redakcije* koji se temelji na glagoljskim tekstovima napisanima na hrvatskome crkvenoslavenskome idiomu i nastalima u razdoblju od 11. do sredine 16. stoljeća. Osim kartoteka s listićima, dio je hrvatskih općih jednojezičnih rječnika nastao kompilacijom prethodnih rječnika, kao *Hrvatski enciklopedijski rječnik* (2002), uz veliku ulogu introspekcije, odnosno oslanjanja na vlastitu jezičnu intuiciju i kompetenciju u poziciji leksikografa, koji je često ujedno i jezikoslovac, a ako nije, tada uz konzultaciju pravopisnih priručnika. U predgovoru *Rječnika hrvatskoga jezika* iz 2000. ne spominju se izvori za rječnik, tek se kaže da je nastao temeljem višegodišnjih priprema i prikupljene građe. Frekvencija se spominje na jednome mjestu: navodi se da se vulgarne, šatrovačke i familijarne riječi obrađuju samo ako su vrlo frekventne, premda se ne objašnjava na koji se način frekventnost utvrdila. U predgovoru Anićeva *Velikog rječnika hrvatskoga jezika* iz 2004. navodi se da je treće izdanje toga rječnika, iz 1998., ocijenjeno kao pouzdan rječnik

hrvatskoga suvremenoga standardnog jezika. Spominje se u predgovoru da je Vladimir Anić započeo raditi na rječniku 1972., samozatajno pišući definicije, prikupljajući regionalizme, žargonizme, frazeologiju, razrađujući koncepciju sve do 1991. kada je u prosincu objavljeno prvo izdanje toga rječnika. Dvadeset je godina taj hrvatski leksikograf i jezikoslovac prikupljao građu za rječnik i neizmjernom nam se štetom čini to što sam rječnik ne ostavlja ni traga o tome koju je građu Anić konzultirao, kakve je stavove o izvorima imao, tek se ponešto može naslutiti iz riječi da Anićeve leksikografske zasade počivaju na namjeri da rječnik ne bude knjiga najboljih riječi, već knjiga svih riječi. U pionire hrvatskih rječnika izrađenih na korpusu pripadaju *Hrvatski čestotni rječnik* (1999) i *Rječnik Marulićeve Judite* (2001). Međutim, hrvatski opći rječnici toga vremena, u trenutku u kojem europska leksikografija kao neizostavni dio rječnika navodi izvore, najčešće korpus ili korpuse na kojima se temelje, ne pridaje tomu više od jednog, vrlo općenitog, retka. Čini se da su stav o nepotrebnosti citata u rječniku čija je uloga regulirati jezika dva stoljeća poslije dijelili i urednici tih rječnika, s obzirom na to da ne sadržavaju citate, ne posvećuju puno mjesta ni pažnje primjerima upotrebe i u predgovorima se ne osvrću na izvore jezičnih podataka na kojima se temelji rječnik.

4.2. Nevolje s korpusom

Danas postoji nekoliko računalnih korpusa za hrvatski jezik: Hrvatski nacionalni korpus (HNK) koji broji 216,8 milijuna pojavnica (Tadić 2009), Hrvatska jezična riznica[11] s 84,9 milijuna pojavnica, mrežni korpus hrWaC s 1,2 milijardi riječi (Ljubešić i Klubička 2014) i Hrvatski korpus govornog jezika s 250 000 pojavnica (Kuvač Kraljević i Hržica 2017). Njihovom pojavom, a zatim i pojavom prvih rječnika koji su ih počeli primjećivati, nazirala se promjena u tom smislu. Rječnici koji među svojim izvorima navode računalne korpuse su *Hrvatski frazeološki rječnik* (2014), *Školski rječnik hrvatskoga jezika* (2015), zatim *Veliki rječnik hrvatskog standardnog jezika* (2015) i nacrt *Mrežnika* u Hudeček i Mihaljević (2017). Uvidom u te rječnike ustanovili smo da se u pravilu radi o *corpus-illustrated* pristupu, što znači da se korpus koristi kao repozitorij primjera koji će oslikati određenu jezičnu pojavu koja i dalje proizlazi iz vlastitog jezičnog znanja pojedinca (v. Geeraerts 2006: 36–37). Za *Hrvatski mrežni rječnik* koji je u izradi navodi se da će se temeljiti na dvama računalnim korpusima hrvatskoga jezika, iz kojih će se izlučiti 10 000 najfrekventnijih natuknica, no lista tih natuknica bit će uspoređena s ručno prikupljenom listom koja je rezultat introspekcije nekolicine leksikografa te će se korigirati prema tome (v. Hudeček i Mihaljević 2017: 176–177). U *Školskome rječniku hrvatskoga jezika* (2015) navodi se da se temelji na računalnome korpusu tekstova Hrvatske jezične riznice te da su obrađivači pri obradi provjeravali značenja i upotrebu riječi u korpusu, ali primjeri nisu izravno preuzimani iz korpusa

[11] URL 1.

nego su ih oblikovali sami obrađivači. U *Hrvatskome frazeološkome rječniku* (2014) navodi se da su autori rječnika, ako nisu u izvorima mogli pronaći potvrdu za frazem, sami ciljano sastavili primjer. Treba spomenuti i da se u predgovoru najnovijeg općeg rječnika hrvatskoga jezika, *Velikoga rječnika hrvatskoga standardnoga jezika*, kao izvori, osim rječnika prethodnika, navode i postojeći računalni jezični korpusi, premda nije detaljnije objašnjeno kako su korpusi korišteni i za što su služili. Kada je riječ o primjerima u tome rječniku kojih na mnogo mjesta uopće nema ili su često šturi i kratki, navodi se da je riječ o „kolokacijama, citatima iz korpusa i dr." (VRH 2015: XII). Iako je taj opis izvora primjera podosta neodređen, posebno zbog kratice *i dr.* za koju nije jasno na što se odnosi, ostaje pretpostaviti da su citati iz korpusa dobiveni metodom *corpus-illustrated*, a da kratica *i dr.* upućuje na ciljano smišljanje ili skraćivanje primjera samih leksikografa.

Može se uočiti da su u hrvatskoj leksikografiji prisutne poteškoće s *prepuštanjem* korpusu, odnosno da se i dalje prednost daje stavovima pojedinca nego dokazima o upotrebi. Kao jedan od razloga često se spominje nereprezentativnost dostupnih korpusa (v. npr. Hudeček i Mihaljević 2018). Naime, korpus koji je zamišljen kao nacionalni, referentni korpus hrvatskoga jezika (HNK) prestao je rasti nakon 216 milijuna pojavnica, pa ga kroatistička struka ne smatra reprezentativnim (v. Hudeček, Mihaljević i Vukojević 2002), kao ni Hrvatsku jezičnu riznicu jer ne sadržava djela novija od 2010. Mrežnome korpusu hrWaC, koji ispunjava uvjet veličine i suvremenosti, kao nedostatak pripisuje se to što sadržava mnogo nestandardnog, područno obilježenog leksika s obzirom na to da uključuje tekstove s domene .hr koji pripadaju većinom publicističkom, razgovornom i administrativnom stilu. Štoviše, taj je korpus podvrgnut provjerama grešaka (v. Blagus Bartolec i Matas Ivanković 2019). Polazeći od stava da jezik koji koriste govornici hrvatskog jezika (a koji se nalazi u hrWaCu) sadržava greške, izdvojene su, osim slovnih, tipografskih i interpunkcijskih pogrešaka, i stilističke pogreške, odnosno 'odstupanja' u tekstovima u kojima autori unose nestandardne oblike i idiomatska jezična obilježja (2019: 38). Riječ je o spajanju nespojivog, o miješanju kategorija: polazni stav o jeziku je preskriptivistički, a koristi se alat korpusne lingvistike kojoj je inherentna deskriptivnost.

Na temelju toga može se zaključiti da se u hrvatskoj leksikografiji izbjegava temeljiti rječnike na korpusu u pravom smislu zato jer se njeguje ideja o rječniku kao mjestu na kojem će se poštivati standard, ali i mjestu koje prema procjeni leksikografa *mora* sadržavati određene riječi, iako se u korpusu ne javljaju često ili uopće. Hrvatski je leksikograf i dalje onaj jezični autoritet čije procjene imaju prednost nad autoritetom jezičnih dokaza. Izvori od kojih se korpus sastoji iz te su perspektive nedovoljno reprezentativni i nedovoljno primjereni. Primjerice, autori *Hrvatskog mrežnog rječnika* (Hudeček i Mihaljević 2018) koji je u izradi smatraju da određene kolokacije nisu karakteristične za pretraženu riječ nego za korpus iz kojeg su preuzete ili su uvredljive, kao npr. *sisata konobarica* koja se u hrWaCu pojavljuje kao prvi rezultat u opciji skice

riječi u alatu Sketch Engine. Najprije treba reći da su rezultati u skicama riječi poredani prema tipičnosti sveza (logDice), što nije isto što i frekvencija, a detaljnijim se uvidom može saznati da je najfrekventnija kolokacija uz *konobaricu zgodna*, zatim *simpatična, ljubazna*, a potom i *sisata* te *brkata*. Dakako, s obzirom na rezultat logDicea, korpusni bi leksikograf svakako trebao u rječnik uključiti i kolokaciju *sisata*, a za problem eventualne uvredljivosti može se pogledati u primjere dobre prakse u drugim leksikografijama, npr. u Macmillanovu rječniku[12] u objašnjenje značenja uključuje se informacija o tome da je riječ o uvredljivu terminu, ili u Longmanovu frazeološkome rječniku (1998) uz frazem daje se napomena: *oprez, neki ljudi ovo mogu smatrati uvredljivim*. Etička pitanja poput političke ili spolne korektnosti u leksikografiji nisu zanemariva, štoviše područje kritičke leksikografije (usp. Moon 2014) posvećuje im mnogo pažnje, međutim, ne možemo se složiti da je prešućivanje rezultata iz korpusa na kojem se neki rječnik načelno temelji dobro rješenje. Osim toga, procjena leksikografa o tome koliko je nešto karakteristično ili tipično za neku riječ, smješta nas u područje introspektivne leksikografije. Intuicija jezikoslovca može biti vrlo vrijedna, ali ako nešto nije, onda nije reprezentativna za jezik u cjelini, što je nedostatak koji se pripisuje najvećem hrvatskom korpusu i ujedno argument za posezanje za vlastitom intuicijom.

Naposljetku, čini se da suvremeni hrvatski opći rječnici nastoje ostati u određenoj sigurnoj sredini. Nijedan naš rječnik zasad nije oblikovan isključivo na korpusu, ali podaci iz korpusa pomalo ulaze u hrvatske rječnike. Vjerujemo da bi rječnici koji sadržavaju riječi i opise na temelju jezične introspekcije jezikoslovaca koji su ujedno i stručnjaci za standardni jezik svakako mogli naći svoje korisnike i zadovoljiti njihove potrebe, no opisom teorijskih postavki suvremene korpusne leksikografije u poglavlju 3.1. ovoga rada željeli smo pokazati da je korpusna leksikografija nešto posve drugo od toga. Smisao korpusnoga pristupa je da leksikograf ne prosuđuje, nego opisuje ono što vidi u korpusu. Pritom se jasno kaže na kojem se korpusu temelji rječnik, a individualne procjene o tome bi li nešto trebalo ući u rječnik, je li ispravno i čini li nam se frekventno tada neizbježno ostaju izvan fokusa.

5. Zaključak: u smjeru funkcionalne leksikografije

Izvori jezičnih podataka na kojima se temelji rječnik važan su dio leksikografskoga rada. U radu je opisano nekoliko vrsta takvih izvora s naglaskom na suvremenu leksikografiju i računalne korpuse. Također, pokazalo se da na leksikografski rad utječu i jezične i izvanjezične okolnosti, a pitanje izvora na kojima se temelji rječnik posebno je upravljano stavovima leksikografa o jeziku. Pogled u hrvatske suvremene opće rječnike pokazao je da se izvorima ne posvećuje previše pažnje, a napomene da su čestotnost i raširenost upotrebe bili važan kriterij za uvrštavanje razgovornih riječi, regionalizama,

[12] URL 2.

posuđenica i tuđica (VRH 2015: 9) ne objašnjavaju se detaljnije, pa ostaje nepoznata metoda utvrđivanja raširenosti upotrebe. Također, rječnici se oslanjaju odnosno usklađuju s pravopisima zadržavajući tako i dalje obilježje propisivanja, a ne opisivanja.

S druge strane, autoru ovih redaka čini se da se na dovoljno drugih mjesta skrbi o normi: uz mobilne aplikacije s jezičnim zadacima, mrežne stranice s jezičnim savjetima, televizijske emisije u kojima se detektiraju jezične pogreške sudionika tih emisija, telefonski broj za jezične savjete, tri pravopisa, doduše neujednačena, govornicima hrvatskog jezika, kao i onima koji taj jezik uče, dobro bi došao rječnik koji će im pokazati kakav je taj jezik u upotrebi, odnosno kako komuniciraju njegovi govornici.

Suvremeni opći rječnik hrvatskog jezika, rasterećen imperativa norme i propisivanja, fokus treba usmjeriti na korisnika i podatke o značenju i upotrebi koje mu može pružiti, na tragu funkcionalne teorije leksikografije. Korisniku toga rječnika važni su i korisni najprije podaci o tome što znači određena riječ: u kojim kontekstima je govornici najčešće koriste, ali i kada je vrlo rijetko koriste, zatim treba moći vidjeti nekoliko primjera stvarne upotrebe u jeziku, a s obzirom na prostornu neograničenost e-rječnika, mogu se dodati i razni drugi podaci, o kolokacijama, frazemima, čestoj ili specifičnoj upotrebi i drugo. Sve te podatke leksikograf temelji na korpusnome istraživanju koje prethodi izradi rječnika. Hrvatske bi korpuse svakako valjalo dopunjavati i osuvremenjivati, ali suština korpusnoga pristupa je, riječima Johna Sinclaira, vjerovati korpusu. I kad napusti svoje procjene o tome što bi *trebalo* biti u rječniku, uloga leksikografa ostaje neizmjerno zahtjevna: uspoređujući rezultate, gledajući primjere treba donositi odluke o tome što i kako reći korisnicima, odnosno kako u rječniku opisati ono što *jest* u korpusu.

Izvori

Anić, Vladimir i sur. [ur.] 2003. *Hrvatski enciklopedijski rječnik.* Zagreb: Novi Liber.

Anić, Vladimir. 2004. *Veliki rječnik hrvatskoga jezika*. Zagreb: Novi Liber.

Benešić, Julije. 1985. *Rječnik hrvatskoga književnog jezika od preporoda do I. G. Kovačića,* sv. 1. Zagreb: JAZU – Globus.

Birtić, Matea i sur. [ur.]. 2015. *Školski rječnik hrvatskoga jezika*. Zagreb: Institut za hrvatski jezik i jezikoslovlje – Školska knjiga.

Jojić, Ljiljana i sur. [ur.]. 2015. *Veliki rječnik hrvatskoga standardnog jezika*. Zagreb: Školska knjiga.

Longman Idioms Dictionary. 1998. Harlow: Longman.

Menac, Antica; Fink Arsovski, Željka; Venturin, Radovan. 2014. *Hrvatski frazeološki rječnik.* Zagreb: Naklada Ljevak.

Šonje, Jure [ur.]. 2000. *Rječnik hrvatskoga jezika*. Zagreb: Školska knjiga.

VRH = 2015. *Veliki rječnik hrvatskoga standardnog jezika* [ur. Jojić, Ljiljana]. Zagreb: Školska knjiga.

Literatura

Atkins, Sue; Rundell, Michael. 2008. *The Oxford guide to practical lexicography*. New York: Oxford University Press.

Barnhart, Clarence Lewis. 1962. Problems in editing comercial monolingual dictionaries. U: *Problems in lexicography* [ur. Householder, Fred; Saporta, Sol]. Bloomington: Indiana University, 161–181.

Baroni, Marco i sur. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43 (3), 209–226.

Béjoint, Henri. 2010. *Lexicography of english*. New York: Oxford University Press.

Bergenholtz, Henning; Nielsen, Sandro; Tarp, Sven [ur.]. 2009. *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang.

Biber, Douglas; Conrad, Susan; Reppen, Randi. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Blagus Bartolec, Goranka; Matas Ivanković, Ivana. 2019. Korpus umom korisnika (na što treba pripaziti u korpusno utemeljenom istraživanju). U: *Jezik i um. Zbornik radova s međunarodnoga znanstvenog skupa Hrvatskoga društva za primijenjenu lingvistiku održanoga od 3. do 5. svibnja 2018.* [ur. Matešić, Mihaela; Vlastelić, Anastazija]. Zagreb: Srednja Europa, 31–42.

Bratanić, Maja. 1991. *Rječnik i kultura*. Zagreb: Filozofski fakultet.

Bratanić, Maja. 1997. Od intuicije do opservacije i natrag (Višejezična leksikografija i paralelni korpusi). *Suvremena lingvistika*, 43/42 (1–2), 1–12.

Cruse, Alan D. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Filipović Petrović, Ivana; Parizoska, Jelena. 2017. Leksikografska obrada frazema s promjenjivom glagolskom sastavnicom u hrvatskome. *Jezikoslovlje*, 18 (2), 245–278.

Filipović Petrović, Ivana. 2018. *Kada se sretnu leksikografija i frazeologija*: *o statusu frazema u rječniku*. Zagreb: Srednja Europa.

Fuertes-Olivera, Pedro A. 2009. The Function Theory of Lexicography and Electronic Dictionaries: WIKTIONARY as a Prototype of Collective Free Multiple-Language Internet Dictionary. U: *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow* [ur. Bergenholtz, Henning; Nielsen, Sandro; Tarp, Sven]. Bern: Peter Lang, 99–134.

Geeraerts, Dirk. 2006. Methodology in Cognitive Linguistics. U: *Cognitive Linguistics: Current Applications and Future Perspectives* [ur. Kristiansen, G. i sur.]. Berlin – Boston: De Gruyter Mouton, 21–50.

Hanks, Patrick. 1990. Evidence and intuition in lexicography. U: *Meaning and lexicography* [ur. Tomaszczyk, Jacek; Lewandowska-Tomaszczyk, Barbara]. Amsterdam, Philadelphia: John Benjamins, 31–41.

Hanks, Patrick. 2012. Corpus evidence and electronic lexicography. U: *Electronic lexicography* [ur. Sylviane Granger; Paquot, Magali]. Oxford: Oxford University Press, 57–82.

Hudeček, Lana; Mihaljević, Milica. 2009. *Hrvatski terminološki priručnik*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Hudeček, Lana; Mihaljević, Milica. 2017. The Croatian Web Dictionary Project – Mrežnik. U: *Electronic lexicography in the 21st century* [ur. Kosem, Iztok]. Leiden: Lexical Computing CZ s.r.o., 172–192.

Hudeček, Lana; Mihaljević, Milica. 2018. Croatian Web Dictionary Mrežnik: One year later – What is different? U: *Proceedings of the Conference on Language Technologies & Digital Humanities* [ur. Fišer, Darja; Pančur, Andrej]. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani, 106–113.

Hudeček, Lana; Mihaljević, Milica; Vukojević, Luka. 2002. Hrvatski nacionalni korpus – kakakv jest i kakav bi trebao biti – kroatistički pogled. U: *Primijenjena lingvistika u Hrvatskoj – izazovi na početku XXI. stoljeća: zbornik* [ur. Ivanetić, Nada; Pritchard, Boris; Stolac, Diana]. Zagreb – Rijeka: Hrvatsko društvo za primijenjenu lingvistiku – Graftrade, 211–221.

Kemmer, Suzanne; Barlow, Michael. 2000. Introduction: A Usage-Based Conception of Language. U: *Usage-Based Models of Language* [ur. Barlow, Michael; Suzanne Kemmer]. Stanford: CSLI Publications, vii–xxviii.

Klobučar Srbić, Iva. 2008. Obol korpusne lingvistike suvremenoj leksikografiji. *Studia Lexicographica*, 2 (3), 39–51.

Kupietz, Marc. 2016. Constructing a corpus. U: *The Oxford handbook of lexicography* [ur. Durkin, Philip]. Oxford: Oxford University Press, 62–75.

Kuvač Kraljević, Jelena; Hržica, Gordana. 2017. Hrvatski korpus govornog jezika (HrAL). *Fluminensia*, 28 (2), 87–102.

Labov, William. 1972. Some principles of linguistic methodology. *Language in Society*, 1(1), 97–120.

Landau, Sydney. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Langacker, Ronald W. 1988. A Usage-Based Model. U: *Topics in Cognitive Linguistics* [ur. Rudzka-Ostyn, Brygida]. Amsterdam – Philadelphia: John Benjamins, 127–164.

Langacker, Ronald W. 2000. A Dynamic Usage-Based Model. U: *Usage-Based Models of Language* [ur. Barlow, Michael; Kemmer, Suzanne]. Stanford: CSLI Publications, 1–64.

Leroyer, Patrick. 2007. Bringing corporate dictionary design into accord with corporate image: From words to messages and back again. U: *Dictionary Visions, Research and Practice* [ur. Gottlieb, Henrik; Mogensen, Jens Erik]. Amsterdam – Philadelphia: John Benjamins, 109–117.

Ljubešić, Nikola; Klubička, Filip. 2014. {bs,hr,sr}WaC – Web Corpora of Bosnian, Croatian and Serbian. U: *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* [ur. Bildhauer, Felix; Schäfer, Roland]. Sweden: Association for Computational Linguistics, 29–35.

McEnery, Tony; Gabrielatos, Costas. 2006. English Corpus Linguistics. U: *The Handbook of English Linguistics* [ur. Aarts, Bas; McMahon, April]. Oxford: Blackwell Publishing Ltd, 33–71.

Meyer, Charles F. 2002. *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.

Moguš, Milan. 2001. *Rječnik Marulićeve Judite*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Moguš, Milan; Bratanić, Maja; Tadić, Marko. 1999. *Hrvatski čestotni rječnik*. Zagreb: Zavod za lingvistiku Filozofskoga fakulteta – Školska knjiga.

Moon, Rosamund. 2008. Dictionaries and collocation. U: *Phraseology: An interdisciplinary perspective* [ur. Granger, Sylviane; Meunier, Fanny]. Amsterdam: John Benjamins Publishing, 313–336.

Moon, Rosamund. 2014. Meanings, ideologies, and learner's dictionaries. U: *Proceedings of the XVI Euralex international congress: the user in focus* [ur. Abel, Andrea; Vettori, Chiara; Ralli, Natascia]. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, 85–105.

Mugglestone, Linda. 2016. Description and prescription in dictionaries. U: *The Oxford handbook of lexicography* [ur. Durkin, Philip]. Oxford: Oxford University Press, 76–93.

Nielsen, Sandro; Mourier, Lise. 2007. Design of a function-based Internet Accounting Dictionary. U: *Dictionary Visions, Research and Practice* [ur. Gottlieb, Henrik; Mogensen, Jens Erik]. Amsterdam – Philadelphia: John Benjamins, 119–135.

Parizoska, Jelena; Filipović Petrović, Ivana. 2016. Uporabni model jezika u leksikografskoj obradi frazeološke varijantnosti u hrvatskome. U: *Metodologija i primjena lingvističkih istraživanja* [ur. Udier, Sanda Lucija; Cergol Kovačević, Kristina]. Zagreb: Srednja Europa, 147–158.

Schubert, Bojana. 2018. Senjani u *Benešićevu rječniku* iliti o lijevoj strani *Rječnika hrvatskoga književnoga jezika od preporoda do I. G. Kovačića*. U: *Zbornik radova 1. senjskog interdisciplinarnog simpozija »Zdravo ste nam, braćo, u kamenu Senju! (nova čitanja)«* [ur. Vukelić, Ana]. Senj: Grad Senj, 111–133.

Schütze, Carson T. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology* (Classics in Linguistics 2). Berlin: Language Science Press.

Sinclair, John. 1984. Lexicography as an academic subject. U: *Lexeter '83 Proceedings* [ur. Hartman, R. R. K. Niemeyer, M.]. Tübingen: Max Niemeyer Verlag, 3–12.

Sinclair, John. 1987. The Dictionary of the Future. *Library Review*, 36, 268–278.

Starčević, Anđel. 2016. Govorimo hrvatski ili 'hrvatski': standardni dijalekt i jezične ideologije u institucionalnom diskursu. *Suvremena lingvistika*, 42 (81), 67–103.

Starčević, Anđel; Kapović, Mate; Sarić, Daliborka. 2019. *Jeziku je svejedno*. Zagreb: Sandorf.

Štrkalj Despot, Kristina; Möhrs, Christine. 2015. Pogled u e-leksikografiju. *Rasprave Instituta za jezik i jezikoslovlje*, 41 (2), 329–353.

Tadić, Marko. 2009. New version of the Croatian National Corpus. U: *After half a century of Slavonic natural language processing* [ur. Hlaváčková, Dana i sur.]. Brno: Masaryk University, 199–205.

Tafra, Branka. 2012. Što je hrvatskoj leksikografiji natuknica? U: *Stručak riječima ispunjen, Zbornik radova posvećen Antici Menac o njezinu 90. rođendanu* [ur. Fink-Arsovski, Željka]. Zagreb: FF press – Filozofski fakultet Sveučilišta u Zagrebu, 111–132.

Tonini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam i Philadelphia: John Benjamins Publishing Company.

Tummers, Jose, Kris Heylen i Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, 1–2, 225–261.

URL 1: http://riznica.ihjj.hr/. Pristup 17. rujna 2020.

URL 2: https://www.macmillandictionary.com/. Pristup 20. rujna 2020.

What is corpus to Croatian lexicography

Contemporary theoretical lexicography as an important feature of the dictionary highlights the reliability of the linguistic evidence on which the dictionary is based. In order to be reliable, evidence should approximate closely to the ways in which people normally use and understand language when engaging in real communicative acts. In other words, reliable evidence is what we learn by observing language in use, preferably on a large number of examples. Evidence comes in several forms, depending on the purpose and type of the dictionary: introspection, the process in which we give an account of a word and its meaning by consulting our own mental lexicon; then previous dictionaries; manually collected evidence from questionnaires and field research or citations from literature and finally, in the digital era, large computer corpora.

In this paper we are dealing with lexicographical evidence on which Croatian contemporary dictionaries are based. More concrete, we are looking into the stand of Croatian lexicographers to raw data for a dictionary. Thereby, we indicate the important role of introspection as evidence in Croatian dictionaries, as well as advantages and disadvantages of currently available corpora for Croatian. Finally, given the above-mentioned aspiration for a reliable dictionary that reflects real use, we will introduce the features which that kind of dictionary should possess, in order to give guidelines to Croatian lexicographers to build one.

Keywords: contemporary lexicography, corpora, general purpose dictionaries, lexicographical evidence