Primjena metoda korpusne lingvistike u jezikoslovnim istraživanjima

Borucinsky, Mirjana

Mirjana Borucinsky

Primjena metoda korpusne lingvistike u jezikoslovnim istraživanjima

uniri

PFRI Pomorski fakultet
Sveučilišta u Rijeci

Mirjana Borucinsky

PRIMJENA METODA KORPUSNE LINGVISTIKE U JEZIKOSLOVNIM ISTRAŽIVANJIMA

Izdavač
Sveučilište u Rijeci
Pomorski fakultet

Autor knjige
dr. sc. Mirjana Borucinsky

Urednica
Vesna Vranić Kauzlarić, dipl. ing.

Recenzenti
dr. sc. Mateusz-Milan Stanojević
dr. sc. Irena Bogunović

Lektura i korektura
Ana Bratulić, mag. educ.

Priprema i tisak
Redak d.o.o. – Split

CIP zapis dostupan u računalnom katalogu
Sveučilišne knjižnice Rijeka pod brojem 150525020.
ISBN 978-953-165-140-0

Prvo izdanje
Odlukom Senata Sveučilišta u Rijeci

KLASA: 602-03/23-03/08
URBROJ: 2170-137-03-23-3
ovo se djelo objavljuje kao izdanje Sveučilišta u Rijeci.

PRIMJENA METODA KORPUSNE LINGVISTIKE U JEZIKOSLOVNIM ISTRAŽIVANJIMA

Mirjana Borucinsky

Pomorski fakultet
Sveučilišta u Rijeci

Rijeka, 2023.

SADRŽAJ

POPIS POKRATA

AF – apsolutna frekvencija (engl. *absolute frequency*)

ARF – srednja reducirana frekvencija (engl. *average reduced frequency*)

BNC – *British National Corpus*

BS – *Korpus brodostrojarstva*

CALL – *Computer-assisted Language Learning*

CHILDES – *The Child Language Data Exchange System*

CLARIN – *Common Language Resources and Technology Infrastructure*

CLASSLA – *CLARIN Knowledge Centre for South Slavic languages*

CNC – *češki nacionalni korpus*

CQL – *corpus query language* (jezik za postavljanje upita korpusu)

CQP – *corpus query programme* (program za postavljanje upita korpusu)

CroDi – *Regensburški dijakronijski korpus hrvatskoga jezika*

CroLTeC – *Croatian Learner Text Corpus* (Učenički korpus hrvatskoga jezika)

CroTag – *Croatian tagger*

DF – frekvencija u dokumentu (engl. *document frequency*)

DGT – *Directorate-General for Translation of the European Commission*

EAGLES – *Expert Advisory Group on Language Engineering Standards*

ENGRI – *Engleske riječi u hrvatskome* (*Korpus hrvatskih internetskih portala* 2014.-2018.)

ERH – *Engleske riječi u hrvatskome*, *ad hoc* korpus

EnTenTen – *the English Web Corpus*

FEATS – *features*

GT – *Google Translate*

HJP – *Hrvatski jezični portal*

HNK – *Hrvatski nacionalni korpus*

Hr500k – korpus za uvježbavanje (engl. *training corpus*) hrvatskoga jezika

hrWaC – *hrvatski mrežni korpus* (hr *Web as Corpus*)

LLC – *The London-Lund Corpus of Spoken English*

LOB – *Lancaster-Oslo/Bergen Corpus*

LSTO – *Lancaster Stats Tool Online*

MI – mjera međusobne informacije (engl. *mutual information*)

MaCoCu – *Massive Collection and Curation of Monolingual and Bilingual Data*

MULTEXT-East – *Multilingual Text Tools and Corpora for Central and Eastern European Languages*

NoSkE – *NoSketch Engine*

NLP – obrada prirodnoga jezika (engl. *natural language processing*)

OPUS – *an open source parallel corpus*

OZJ – općeznanstveni jezik

POS – *part of speech tagging*

RAPUT – računalni asistent za pomoć pri unosu teksta osobama s jezičnim poremećajima

Regex – regularni izraz (engl. *regular expression*)

RELDI – *Regional Linguistic Data Initiative*

RF – relativna frekvencija (engl. *relative frequency*)

Riznica – *Hrvatska jezična riznica*

SARGADA – sintaktička i semantička analiza dopuna i dodataka u hrvatskom jeziku

SemCRo – *semantic hypergraph corpus*

SEU – *Survey of English Usage*

SkE – *Sketch Engine*

TTR – *type/token ratio* (omjer pojavnica i različnica)

UD – *universal dependencies* (univerzalne ovisnosti)

ukWaC – mrežni korpus britanskoga engleskog jezika (uk *Web as Corpus*)

UPOS – *universal POS*

XPOS – *treebank-specific POS tagset*

POPIS SLIKA

POPIS TABLICA

UVODNA RIJEČ

Temeljno je polazište ove knjige da je korpusna lingvistika metodološki pristup koji omogućuje empirijsko istraživanje jezika i jezičnih varijeteta a rezultati dobiveni korpusnim metodama imaju veću valjanost od onih do kojih se došlo intuicijom, refleksijom ili promišljanjem. Poznavanje teorijskih postavki korpusne lingvistike, kao i mogućnosti i funkcionalnosti postojećih računalnojezikoslovnih resursa i alata, korisnicima će olakšati interpretaciju podataka dobivenih iz korpusa te rasvijetliti jezične pojave koje istražuju.

Naslanjajući se na Stefanowitscha (2020: 1) korpusnu lingvistiku smatramo vrstom lingvističkoga propitivanja koje se temelji na podacima dobivenima iz korpusa (Stefanowitsch 2020: 1). Drugim riječima, korpusna lingvistika je način pronalaženja odgovora na jezikoslovna istraživačka pitanja utemeljena na cjelovitoj i sustavnoj analizi načina na koji su jezične pojave distribuirane u jezičnome korpusu (Stefanowitsch 2020:55, prev. a.).[1]

Motivacija za radom na korpusima nastala je prilikom izrade doktorske disertacije (Borucinsky, 2015) u kojoj su ispitivane sintaktičke strukture u tadašnjemu jedinom slobodno dostupnom korpusu hrvatskoga općeg jezika – *Hrvatskome nacionalnom korpusu* (*HNK*). Sintaktičke strukture predstavljaju jedan od najvećih izazova u korpusnim istraživanjima pa se korisnik, unatoč dostignućima korpusne lingvistike i razvoju jezičnih tehnologija, često nađe pred zidom prilikom istraživanja jer mu je potrebna funkcionalnost koju postojeći alati ne pružaju. Dok se većina jezikoslovaca slaže oko načina lematizacije i označavanja vrsta riječi, označavanje sintaktičkih kategorija u pravilu polazi od neke teorije, što znači da je teško postići konsenzus oko označavanja sintaktičkih kategorija. Nadalje, u korpusu pronalazimo tipične i karakteristične jezične obrasce pa se jezikoslovci prilikom pretrage manje tipičnih obrazaca, kao što su, primjerice, sintaktičke kategorije (npr. imenska skupina) ili pak engleske ili strane riječi u korpusu hrvatskoga jezika, moraju dovinuti različitim načinima kako da dobiju valjane podatke iz korpusa. Nadalje, jezikoslovci u pravilu nisu programeri, pa im je stoga potrebna podrška informatičara u razvoju resursa. Ova knjiga napisana je iz perspektive korisnika korpusa u nadi

[1] U izvorniku: „Corpus linguistics is the investigation of linguistic research questions based on the complete and systematic analysis of the distribution of linguistic phenomena in a linguistic corpus” (Stefanowitsch 2020: 55).

da će postavljena pitanja i kritički osvrt na mogućnosti postojećih alata doprinijeti razvoju jezičnih tehnologija za hrvatski jezik.

U knjizi se, dakle, preispituju mogućnosti dostupnih računalnojezikoslovnih alata i resursa za hrvatski jezik, koji pripada jezicima sa slabije razvijenim jezičnim tehnologijama. S obzirom na to da korpusi omogućuju uvid u jezik kojemu se ne može pristupiti izravno, već samo preko njegova ostvaraja, postavlja se i pitanje mogu li se korpusni podaci adekvatno ponoviti i mjeriti. Ovaj problem preispituje se iz nekoliko perspektiva. Knjiga je prvenstveno zamišljena kao studija prednosti i nedostataka metoda korpusne lingvistike u obradi hrvatskoga jezika. Uz to se ukazuje i na potrebu za interdisciplinarnim pristupom proučavanju jezika te daljnjim obrazovanjem i usavršavanjem jezikoslovaca u području korpusne lingvistike.

Korpusna lingvistika kao disciplina nije zastupljena u Hrvatskoj, i na mnogim se sveučilištima u Hrvatskoj ne poučava na studijima filoloških usmjerenja, primjerice, na Sveučilištu u Rijeci, iako je potreba za takvim kolegijem prepoznata. Iskorak u tom smislu učinio je B. Perak razvojem mikrokvalifikacije *Jezične tehnologije i digitalna obrada teksta* koja je zamišljena kao strukturirani program koji uključuje ponudu niza kolegija različitih sastavnica Sveučilišta u Rijeci i njihovih studijskih programa, između ostaloga, i kolegij *Korpusna lingvistika*.

Kao pokazatelj da korpusna lingvistika nije rasprostranjena u Hrvatskoj možemo navesti jedan diplomski rad u kojemu se, između ostaloga, raspravlja o položaju korpusne lingvistike u Hrvatskoj. Naime, Hasanić (2017) je istražio položaj korpusne lingvistike među studentima preddiplomskoga, diplomskoga i poslijediplomskoga studija filologije, te nekoliko doktora znanosti, i onima koji su stekli neki stupanj obrazovanja u lingvistici, posebno se osvrnuvši na prednosti korpusne lingvistike kao istraživačkoga alata za proučavanje vokabulara, jezičnih uzoraka i obrazaca te semantičke interpretacije i gramatičkoga aspekta jezika. U tu svrhu proveo je anketno istraživanje u kojem je sudjelovalo 100 ispitanika kako bi utvrdio u kojoj su mjeri upoznati s korpusima i njihovom uporabom. Rezultati su pokazali da 60 % ispitanika nije imalo doticaj s niti jednim oblikom organizirane izobrazbe o korpusnoj lingvistici a 80 % ih nikada nije odslušalo kolegij iz korpusne lingvistike, no gotovo 60 % njih služi se korpusnim alatima i metodama u obrazovanju ili na radnome mjestu. Najveći postotak, točnije 60 % ispitanika koji se služe korpusnim alatima i metodama spada u dobnu skupinu od 20 do 29 godina.

Ovo istraživanje provedeno je davno i na malom uzorku koji nije reprezentativan, no ipak je svojevrsni pokazatelj nezastupljenosti korpusne lingvistike u Hrvatskoj.

Stoga ova knjiga, iako nema primarnu pedagošku namjenu, može poslužiti kao literatura na kolegiju koji se bavi korpusnom lingvistikom, odnosno može biti korisna studentima diplomskih i poslijediplomskih studija filološkoga usmjerenja.

U knjizi su prikazana istraživanja koje sam samostalno ili u suautorstvu provela i objavila, a tiču se načina sastavljanja korpusa te dohvata i interpretacije podataka dobivenih iz korpusa na različitim jezičnim razinama.

Knjiga je podijeljena na sedam poglavlja. U prvome su poglavlju, uz kratki povijesni pregled razvoja korpusne lingvistike, postavljene teorijske osnove korpusa, donosi se definicija korpusa, razmatraju problemi uravnoteženosti, reprezentativnosti i veličine korpusa. Uz opis različitih vrsta korpusa, naglasak se stavlja na mrežne korpuse te se daje iscrpan popis postojećih računalnojezikoslovnih resursa i alata za hrvatski jezik. U ovome se poglavlju opisuje i kako pristupiti korpusu te kako ga pretraživati. Drugo poglavlje posvećeno je strojnome obilježavanju korpusa, kao preduvjetu za ispitivanja jezika. Neobilježeni korpus koji ne sadrži dodatne jezične podatke nema veliku vrijednost, stoga je u ovome dijelu najbitnija suradnja jezikoslovaca i programera, ili inženjera koji razvijaju sustave. Treće poglavlje donosi terminologiju koja je potrebna za razumijevanje korpusa i pozadinskih procesa koji se događaju kada računalu damo naredbu poput „pronađi sve opće imenice u korpusu općega hrvatskog jezika“. Osnovni pojmovi ilustrirani su formulama i primjerima. Četvrto poglavlje prikazuje načine sastavljanja ili izrade vlastitoga korpusa uz napomenu o tome koje odluke valja donijeti pri izradi korpusa te uz raspravu o ograničenjima s kojima se istraživači susreću pri izradi korpusa. Posebno je opisana izrada korpusa za pronalaženje engleskih riječi u hrvatskome te specijaliziranih korpusa. Peto i šesto poglavlje odnose se na dohvaćanje i interpretaciju podataka iz korpusa kroz evidenciju i frekvenciju. U šestom se poglavlju bavimo tematikom crpljenja engleskih riječi iz korpusa hrvatskoga jezika. U posljednjem, sedmom poglavlju preispituje se treća vrsta podataka koja se može dobiti iz korpusa, odnosno relacija. Pri tome se istraživanja ograničavaju na razinu riječi, sintagme (skupine) i rečenice te se preispituje suodnos leksika i gramatike iz perspektive korpusne lingvistike.

1. KORPUSNA LINGVISTIKA

U prvome se poglavlju, uz kratki povijesni pregled razvoja korpusne lingvistike, postavljaju teorijske osnove korpusa kao metodološkoga konstrukta, daje se definicija korpusa te razmatraju problemi uravnoteženosti i veličine korpusa. Uz opis različitih vrsta korpusa, naglasak se stavlja na mrežne korpuse te se daje opis razvoja jezičnih tehnologija za hrvatski jezik te iscrpan popis postojećih računalnojezikoslovnih resursa i alata za hrvatski jezik. Na kraju ovoga poglavlja opisan je način pristupa korpusu i pretraživanja korpusa.

1.1. POVIJESNI PREGLED

Na samom početku poslužit ćemo se metaforom o kapljici vode u moru (Březina, 2021) kojom možemo predočiti pojam korpusa. Naime, korpus kao jezični uzorak predstavlja kap vode u beskrajnome moru jezične proizvodnje kojoj svakodnevno svjedočimo. To prije svega vrijedi za opći jezik jer se procjenjuje da osoba u prosjeku izgovori oko 16 000 riječi dnevno (Březina, 2018: 15). Pomnožimo li taj broj s brojem govornika hrvatskoga jezika[2], dobit ćemo 880 milijardi riječi koje se dnevno napišu ili izgovore. S obzirom na količinu teksta koja se svakodnevno piše i riječi koje se izgovaraju, nemoguće bi bilo proučiti sve pojavnosti, stoga se, baš kao u prirodnim znanostima u kojima se, primjerice, pomoću epruvete uzima uzorak kojim se ispituje razina saliniteta mora, u korpusnoj lingvistici uzima reprezentativan jezični uzorak na temelju kojega proučavamo određenu jezičnu pojavu. Kao što oceanolozi i biolozi ne ispituju cijelo Jadransko more da bi utvrdili njegovu kakvoću, tako ni jezikoslovci ne proučavaju sveukupnu jezičnu proizvodnju, već odabiru reprezentativan uzorak koji će podvrgnuti analizi. Taj je uzorak korpus, na kojemu možemo mjeriti učestalost pojavljivanja jezičnih obrazaca, njihove značajke, varijetet ili funkcionalni stil u kojem se najčešće pojavljuju, ili načine na koje se s vremenom mijenjaju. Tako se, primjerice, može sastaviti nekoliko malih uzoraka od po milijun riječi koje čine neformalni razgovori, novinski članci, predavanja, isječci iz književnih djela itd., u kojima se može pratiti negativan trend uporabe modalnoga glagola *must* ‘morati’ u engleskome kao indikaciju da engleski u današnje vrijeme daje prednost manje izravnome načinu obraćanja

[2] Izvor: *O Hrvatskome Jeziku - Institut Za Hrvatski Jezik i Jezikoslovlje*.

(Březina, 2021). Ujedno takvi trendovi mogu odražavati društvenu stvarnost, odnosno mogu ukazati na promjene u društvenoj strukturi.

Prvi značajniji projekt prikupljanja podataka za empirijsko proučavanje gramatike pod nazivom *Survey of English Usage* (*SEU*)[3] pokrenuo je R. Quirk 1959. godine. Projekt je predstavljao polazište za empirijsko proučavanje jezika (Teubert i Čermáková, 2007) i svojevrsna je prekretnica u razvoju moderne korpusne lingvistike. Osim Sveučilišta u Londonu i Birminghamu, za razvoj korpusne lingvistike važno je i Sveučilište Lancaster gdje je započet rad na razvoju prvoga velikog računalnog korpusa pisanoga britanskog engleskog jezika, a taj je projekt kasnije iz financijskih razloga nastavljen u Norveškoj i danas poznat kao *Lancaster-Oslo-Bergen Corpus*[4] (*LOB*). I dandanas sveučilište Lancaster jedno je od vodećih sveučilišta iz područja korpusne lingvistike na kojem se okupljaju (ili su se okupljala) imena poput G. Leecha, T. McEneryja, A. Hardieja, V. Brezine.

Među najvažnije resurse u prvim fazama razvoja korpusne lingvistike svakako valja ubrojiti i publikaciju *Computational Analysis of Present-Day American English* (Kucera i Francis, 1970), utemeljenu na analizi milijunskoga korpusa *Brown* (engl. *Brown Corpus*). Začetke korpusne lingvistike, dakle, nalazimo u 60-im godinama 20. stoljeća[5]. Na početku svog razvoja korpusna lingvistika bila je marginalizirani pristup koji se pojavio otprilike u vrijeme kada je Chomsky[6] uveo velike promjene u lingvistiku. U tom se razdoblju korpusna lingvistika uglavnom koristila kao pomoć pri istraživanju engleskoga jezika, osobito gramatike i leksikona, te pri učenju i poučavanju engleskoga kao stranoga (drugoga) jezika. No, veliki uspon korpusne lingvistike dogodio se 90-ih godina 20. stoljeća kada su osobna računala postala dostupnija. Posljednja dva desetljeća 20. stoljeća obilježena su „punim cvatom korpusne lingvistike, i njezinim prerastanjem iz metodologije u nov pristup promatranju jezične građe iz kojega proizlazi i nova vrsta znanja o jeziku“ (Bratanić, 1998: 172, usp. također Leech, 1992). Od 90-ih godina pa sve do danas korpusna je lingvistika

[3] Taj korpus kasnije je postao poznat pod nazivom *The London-Lund Corpus of Spoken English (LLC)*, a predstavlja prvi strojno čitljiv korpus govornoga jezika koji se sastoji od 200 tekstova i oko 5 000 riječi unutar svakog teksta.

[4] Korpus *LOB* sadržavao je 1 milijun riječi, što je za ono doba bio veliki poduhvat.

[5] O razvoju korpusne lingvistike u Hrvatskoj v. 1.7.

[6] Ranih 60-ih godina 20. stoljeća Chomskyjevi radovi pokrenuli su svojevrsnu revoluciju u lingvistici te žarište istraživanja jezika preusmjerili s empirizma i jezične uporabe (engl. *performance*) k racionalizmu i jezičnoj sposobnosti (engl. *competence*).

postala nezaobilazna u jezikoslovnim istraživanjima, a zanimanje za korpuse ne javlja se samo među jezikoslovcima već i u medijima, području obrazovanja i informacijskih znanosti koje preuzimaju tehnike korpusne lingvistike u proučavanju jezičnih obrazaca ili strukture jezika. Korpusna lingvistika korisna je svakome koga zanimaju velike zbirke tekstova, način pristupanja istima te mogućnosti njihove analize.

1.2. KORPUSNA LINGVISTIKA KAO TEORIJA I METODOLOGIJA

Bekavac (2001: 1) postulira da korpusna lingvistika označava istraživanje jezika na temelju korpusa, no Leech (2011) upozorava da naziv korpusna lingvistika[7] nije najprimjereniji, jer korpusna lingvistika nije grana lingvistike u tradicionalnome smislu, kao što je to primjerice sociolingvistika, već je put k lingvistici.[8] Korpuse i korpusnu lingvistiku kao metodologiju često se poistovjećuje s izumom mikroskopa ili Hubbova teleskopa jer omogućuje uvid u jezik, odnosno da jeziku pristupimo preko njegova ostvaraja. Najveća prednost korpusne lingvistike očituje se u mogućnosti istraživanja uporabe jezika, kao što sljedeća, često citirana rečenica potvrđuje: „Korpusni lingvisti proučavaju stvarni jezik, ostali lingvisti sjede za stolom i razmišljaju o divljim, nemogućim rečenicama“ (McEnery i Wilson, 2001: 1, prev. a.).[9]

Kako Leech (2011) daje naslutiti, ne postoji suglasnost oko statusa korpusne lingvistike i njezine precizne definicije. Dok je jedni smatraju zasebnom teorijskom disciplinom (npr. Leech 1992: 106; Stubbs 1993: 2f; Teubert 2005: 2; Tognini-Bonelli 2001: 1), drugi je smatraju metodologijom[10] (npr. Bowker i Pearson, 2002; Gries, 2010; Hardie i McEnery, 2010; McEnery i dr., 2006 i dr.).[11] S obzirom na to da je korpusna lingvistika kao metodološki alat prožeta kroz sve jezične discipline (od fonetike do diskursa), s pravom je Mukherjee

[7] Smatra se da je naziv *korpusna lingvistika* (engl. *corpus linguistics*) uveo J. Aarts 1980. god. (usp. Leech, 2011).

[8] U izvorniku: “Corpus linguistics is not a branch of linguistics, but the route into linguistics” (Hoey, 1998, citirano u McCarthy, 2004).

[9] U izvorniku: „Corpus linguists study real language, other linguists just sit at their coffee table and think of wild and impossible sentences” (McEnery i Wilson 2001: 1).

[10] U engleskome govornom području ove dvije struje poznate su kao *corpus as theory* (*school of Birmingham*) i *corpus as method* (*school of Lancaster*) (McEnery i Hardie, 2012).

[11] Usp. također Lalli Paćelat (2014).

(2010) propitivao njezin status,[12] tj. može li korpusna lingvistika i dalje postojati kao zasebna teorijska disciplina i u kojem će se smjeru razvijati. Gries (2012: 43)[13] smatra da korpusna lingvistika nije lingvistička teorija, kao što to, primjerice, nije ni eksperimentalna lingvistika koja također preispituje jedinice, strukture i procese unutar formalne lingvistike.

Zaključno, korpusnu lingvistiku smatramo metodološkim pristupom koji se primjenjuje u jezičnoj analizi koristeći pritom napretke računalne tehnologije i oslanjajući se na stručnost jezikoslovaca u analizi uporabe jezika. Bez uporabe računala teško bismo sustavno mogli proučavati jezik, a u konačnici bi takvo proučavanje bilo dugotrajno, neučinkovito i nepouzdano, posebice uzmemo li u obzir ljudski faktor i mogućnost pogrešaka pri ručnoj analizi velike količine podataka. Korpusna se lingvistika koristi empirijskim dokazima dostupnima u korpusima kako bi dala uvid u to kako se jezik koristi u različitim društvenim situacijama.

1.3. KORPUSNI PRISTUPI JEZIKOSLOVNIM ISTRAŽIVANJIMA

Korpus je velika zbirka izvornih tekstova, tj. jezičnih uzoraka koji su nastali u stvarnim komunikacijskim situacijama (Stefanowitsch 2020: 1). Korpusu kao jezičnome uzorku možemo pristupiti na dva načina (Tognini-Bonelli, 2001):

1. Pristup utemeljen na korpusu (engl. *corpus-based approach*)
2. Pristup vođen korpusom (engl. *corpus-driven approach*).

U prvome se pristupu provjeravaju unaprijed postavljene hipoteze (npr. jezik struke sadržavat će veći broj višerječnih naziva u odnosu na opći jezik). Riječ je o deduktivnom ili tzv. *top down* pristupu jeziku u kojemu se korpus promatra iz perspektive postojeće teorije (npr. iz perspektive sistemske funkcionalne gramatike u korpusu se traže jezični obrasci koji odgovaraju opisu imenske skupine unutar te teorije, usp. Borucinsky, 2015). Kao takav, pristup utemeljen na korpusu ima iste ciljeve kao i funkcionalna lingvistika, tj. da opiše i objasni jezične obrasce te njihovu uporabu i varijaciju. Istraživanja temeljena na ovome pristupu pokazala su da se jezična obilježja razlikuju prema funkcionalnim stilovima, odnosno da svaki stil raspolaže određenim jezičnim obrascima čime

[12] Detaljnije v. Lalli Paćelat (2014).

[13] Gries (2012) ispituje odnos između korpusne lingvistike te kognitivne lingvistike i psiholingvistike.

se propituju opći lingvistički opisi jezika i pokazuje da je opis jezika koji ne uzima u obzir funkcionalne stilove nepotpun. Najveće su prednosti ovog pristupa pouzdanost i valjanost (usp. Biber 2015: 197). Najveći je nedostatak pristupa utemeljenoga na korpusu fokusiranost na jezičnu pojavu koja se istražuje, što može dovesti do toga da istraživaču promaknu pojave izvan onoga što istražuje.

Za razliku od toga, u pristupu koji je vođen korpusom hipoteze se postavljaju na temelju rezultata korpusne analize. U korpusima koji se koriste u potonjem pristupu gramatičke kategorije i jezični obrasci proizlaze iz analize korpusa, a takva analiza podrazumijeva tek postojanje riječi, dok gramatičke kategorije i sintaktičke strukture nisu unaprijed definirane prema nekoj lingvističkoj teoriji, odnosno nemaju, kako navodi Biber (2015: 196) *a priori* status pa jezični opis proizlazi iz jezičnih obrazaca koji su uočeni u korpusu. To je tzv. induktivni ili *bottom up* pristup, jer novi jezični obrasci proizlaze iz korpusnih podataka (Biber, 2009: 276; Gries, 2009: 328), a ponajprije se primjenjuje u leksikografskim istraživanjima.

Primjer rječnika koji je nastao na temelju korpusom vođenoga pristupa jest *Hrvatski čestotni rječnik* (Moguš, Bratanić i Tadić, 1999), dok je *Hrvatski mrežni rječnik,* prema riječima autora, „korpusno utemeljen[14] rječnik u kojemu se obrađivač služi korpusom, ali može slobodno procijeniti što treba unijeti u rječnik te rječnik može po potrebi dopunjavati i riječima iz drugih izvora te kolokacijama i značenjima koji nisu potvrđeni u korpusu".[15] Međutim, ovdje valja spomenuti još jedan pristup - tzv. korpusno oprimjeren pristup (engl. *corpus-illustrated approach*), (Filipović Petrović, 2018; Tummers i dr., 2005) u kojemu se polazi od intuicije te se potvrde traže u korpusu, a ako ih nema u korpusu sastavljači ili priređivači rječnika ih smisle. Time se zamjenjuju introspektivno izmišljeni podaci ili primjeri uporabe s introspektivno izabranim podacima sa svim nedostacima, kao što je ne uzimanje u obzir čestotnih lista i sl. Ovaj pristup bolje opisuje način na koji je sastavljen *Hrvatski mrežni rječnik*, no on nije dijelom korpusne lingvistike jer se podaci ne prikupljaju sustavno, a potom podvrgavaju interpretaciji.

Rasprave se vode o tome može li pristup u potpunosti biti vođen korpusom, odnosno je li moguće promatrati tekst na sasvim neutralan način. Gries (2012),

[14] O korpusno utemeljenim rječnicima u hrvatskome v. Štrkalj Despot i Ostroški Anić (2020).

[15] *Pojmovnik - Hrvatski mrežni rječnik*.

primjerice, smatra da ne postoji pristup koji je u potpunosti vođen korpusom, a kao iznimku navodi tzv. *Linear Unit Grammar* (McHardy Sinclair i Mauranen, 2006). Istraživanja vođena korpusom uključuju i elemente istraživanja temeljenih na korpusu, te se Biber (2009) stoga zalaže za hibridni pristup.

Ova dva pristupa ujedno i odražavaju neodređenost statusa korpusne lingvistike kao discipline ili metodologije (v. 1.1., usp. također Lalli Paćelat (2014: 57)). Također, nije moguće tvrditi koji je od dvaju pristupa bolji, jer odabir pristupa ovisi o tome što se istražuje.

Smatramo da je teško postići potpunu objektivnost i promatrati korpusne podatke, a da ne polazimo od neke teorije ili određene jezične kategorije, pa su stoga istraživanja koja su prikazana u ovoj knjizi utemeljena na korpusu.

1.4. DEFINICIJA KORPUSA

Korpus[16] se definira kao velika zbirka tekstova tj. riječi u kontekstu ili potpunih rečenica s dodatnim informacijama o samome tekstu. Sinclair (1991: 171, prev. a.) navodi da je korpus zbirka dijelova teksta u elektroničkome obliku koji su odabrani prema vanjskim kriterijima tako da predstavljaju, koliko je to moguće, jezik ili jezični varijetet, a služe kao izvor podataka za jezikoslovna istraživanja.[17] Uz definiciju korpusa tradicionalno se vežu četiri ključne riječi (v. također McEnery i dr., 2006):

1. strojna čitljivost
2. prirodnost/izvornost

[16] Engleska riječ *corpus* potječe od latinske riječi *corpus* koja je po prvi puta zabilježena u uporabi u 13. stoljeću. Riječ se 90-ih godina 20. stoljeća počela rabiti u značenju zbirke tekstova koja predstavlja uzorak određenoga jezičnog varijeteta (Izvor: *Dictionary.Com | Meanings and Definitions of Words at Dictionary.Com)*. Hrvatska riječ *korpus* nije potvrđena u Rječniku hrvatskoga ili srpskoga jezika, poznatijim pod nazivom Akademijin rječnik, što znači da je nastala kasnije. Riječ korpus *Hrvatski jezični portal* (HJP, 2022) definira kao „cjelovitu zbirku podataka, dokumenata, građe za neku disciplinu (npr. korpus riječi hrvatskoga jezika, korpus srednjovjekovnih dokumenata)*“,* dok je u *Hrvatskoj enciklopediji* (HE, 2022) natuknica obrađena na sljedeći način: *korpus (lat. corpus: tijelo) „3. U jezikoslovlju, ukupnost izričaja koji se podvrgavaju analizi kako bi se opisao dani jezik (njegov glasovni, gramatički i dr. ustroj). Kako nema izgleda da bi se mogli prikupiti svi izričaji koje su za trajanja ispitivanja ostvarili svi članovi određene jezične zajednice, korpus kao završen skup pisanih ili snimljenih (i transkribiranih) izričaja, tekstova, nužno je praktično rješenje, pa zato treba nastojati da on bude »reprezentativan«. Opisivača ništa ne sprječava da posegne za relevantnom građom i izvan korpusa. Metoda korpusa nameće se osobito onda kada kontakti s ispitanicima traju ograničeno, ili kada se ne mogu obnoviti.“*

[17] U izvorniku: „A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research” (Sinclair, 1991: 171).

3. uravnoteženost uzorka
4. reprezentativnost.

Da bismo zbirku tekstova mogli nazvati korpusom, ona mora biti strojno čitljiva i uključivati tekstove u pisanome i/ili govornome obliku koji su nastali u prirodnim komunikacijskim situacijama, što isključuje metajezik. Korpus kao uravnoteženi uzorak podrazumijeva primjenu skupa jasnih kriterija prilikom odabira tekstova, a odluke se odnose na to hoće li se, primjerice, tekstovi uključivati u korpus u potpunosti ili će se nasumično odabrati uzorci određene duljine, kako će se uzorci stilski i tematski podijeliti, hoće li se i u kojem obujmu obuhvatiti i govoreni jezik ili će se sastaviti korpus isključivo pisanoga jezika itd. (McEnery i Hardie, 2012: 8–11, usp. također Mikelenić, 2020). Nadalje, postoje i korpusi koje McEnery i Hardie (2012: 11) nazivaju oportunističkim korpusima (engl. *opportunistic corpus*), a odnose se pretežno na manje korpuse koji se iz različitih razloga kao što su dostupnost tekstova, tehničke poteškoće i sl. nisu mogli prikupiti kao uravnoteženi korpusi, nego su se u njih uključili svi tekstovi koje je u danom trenutku bilo moguće prikupiti (usp. Mikelenić 2020: 172). Biber (1993) smatra da korpus mora biti dovoljno velik da bude reprezentativan za jezičnu pojavu koja se istražuje. Jedno je od ključnih pitanja vezanih uz korpus i njegova reprezentativnost.[18]

Zanettin (2014) smatra da reprezentativnost valja shvatiti kao relativnu, a ne kao apsolutnu kategoriju, a upitno je i u kojoj mjeri ju je moguće postići. Nadalje, reprezentativnost je lakše postići u manjim, specijaliziranim korpusima, a određeni žanrovi ne zahtijevaju velik korpus, budući da se dodavanjem novih elemenata u korpus ne povećavaju leksikogramatičke i stilske varijacije, odnosno broj različnica ne povećava se srazmjerno broju riječi (Corpas i Seghiri, 2009; McEnery i Wilson, 2001). Pojam reprezentativnosti je zastario, pa se u nastavku knjige koristi naziv *uravnoteženost* kada se opisuje način na koji korpus kao jezični uzorak odražava jezik u ključnim aspektima relevantnima za istraživačka pitanja.

Stefanowitsch (2020: 35) smatra da je s praktičnoga (a možda i s teorijskoga aspekta) nemoguće sastaviti reprezentativan korpus te da se stoga valja voditi obilježjem raznovrsnosti (engl. *diversity*) kao najboljom mogućom zamjenom za reprezentativnost. Da bi korpus bio raznovrstan (ili uravnotežen) u njemu kvantitativno i kvalitativno moraju biti zastupljeni tekstovi jezičnoga

[18] Seghiri (2012) je razvila alat za mjerenje reprezentativnosti korpusa.

varijeteta koji pronalazimo u govornoj zajednici čiji je jezik zastupljen u korpusu (Stefanowitsch 2020: 29). Drugim riječima, time se označava da je uzorak podskup veće populacije tako da uzorak i populacija imaju istu distribuciju.

U praksi se uravnoteženost pri izradi vlastitoga korpusa sastoji od nekoliko ključnih odluka (v. 4.). Među prvima je dizajn korpusa, pri čemu odlučujemo koje vrste tekstova želimo uključiti u uzorak i u kojim omjerima, kako bismo u konačnici dobili uravnotežen korpus. Na potonje utječu tradicija poimanja sličnih korpusa te usporedba vlastitoga korpusa s postojećima, kao i specifična istraživačka pitanja i šire potrebe istraživačke zajednice s kojom želimo podijeliti korpus. Nakon prve faze slijedi prikupljanje podataka relevantnih za istraživanje metodom stratificiranoga nasumičnog uzorkovanja, gdje je to moguće, kako bi se izbjegla pristranost i povećala reprezentativnost korpusa. Koliko je korpus dobar ovisi o tome što istražujemo. Drugim riječima, korpus je dovoljno dobar ako nam omogućuje da istražujemo ono što želimo. Tu se ističe bitna značajka - korisnici i sastavljači korpusa ujedno su i njegovi evaluatori. Naime, korpus može pokazati samo one podatke koje sadrži, a Leech (1992) upozorava da se u veličini korpusa kriju moguće zamke jer veličina korpusa ne mora biti razmjerna raznovrsnosti (i uravnoteženosti) građe (v. također Bratanić 1998: 173). Naime, korpus predstavlja samo tekstove zastupljene u njemu, a ne lingvistički univerzum (Teubert, 1995: 119).

Uz četiri ključne riječi koje su navedene na početku ovoga poglavlja, jedno je od bitnih pitanja pri izradi korpusa i njegova veličina. U provođenju istraživanja često se pitamo je li korpus dovoljno velik za istraživanje jezične pojave kojom se bavimo. Načelno vrijedi pravilo: „Što veće, to bolje“, no odgovor na pitanje ipak ovisi o tome što istražujemo. Korpus je dovoljno velik ako sadrži dovoljan broj pojava koje istražujemo. Primjerice, istražujemo li jezik u djelima Marka Marulića, logično je da takav korpus ne može biti velik, posebice u usporedbi s mrežnim korpusima. Na početku razvoja korpusa i korpusne lingvistike veliki korpusi sadržavali su oko milijun riječi (primjerice korpus *Brown*), a danas govorimo o milijardama riječi (*hrWaC* primjerice sadrži 1,2 milijarde riječi, dok mrežni korpus engleskoga jezika (*EnTenTen 20*) sadrži 38,15 milijardi riječi), iz čega slijedi da je najočitija tendencija u razvoju korpusa rast njihove veličine (Bratanić 1998: 173). Mali korpusi u pravilu su dostatni za proučavanje čestih gramatičkih pojava kao što su, primjerice, genitiv ili akuzativ, no nisu dostatni za proučavanje rjeđih pojava kao što je vokativ. Korpusi srednje veličine prigodni su za istraživanje konkretnih frekventnih riječi kao što su najčešće

imenice u hrvatskome[19], dok su veliki korpusi pogodni za proučavanje manje čestih pojava kao što su specifični frazemi, sinonimi, antonimi i sl. (v. 7.1) te za proučavanje sintaktičkih obrazaca (v. 7.2.).

Zaključit ćemo ovu raspravu oslanjajući se na Anthonyjev (2019b) stav da vrijednost korpusa nije u njegovoj veličini, već u podacima koji se iz njega mogu dobiti. Također, Bowker i Pearson (2002) navode da se više podataka može dobiti iz specijaliziranoga korpusa koji je manji, uravnotežen i pažljivo sastavljen, nego iz većeg korpusa koji nije prilagođen posebnim potrebama istraživanja. No, korpus ne smije biti premalen jer neće sadržavati pojmove ili strukture važne za istraživanje određenog segmenta jezika, pa se stoga neće moći donijeti valjani zaključci.

1.5. VRSTE KORPUSA

Postoji niz podjela korpusa prema vrstama (usp. primjerice Precht i dr., 1998; Sinclair, 1991; Tognini Bonelli 2001; Tognini-Bonelli i Sinclair, 2006), no u ovome potpoglavlju prikazat će se sažeta i pojednostavljena podjela s posebnim osvrtom na mrežne korpuse koji su relevantni za istraživanja prikazana u ovoj knjizi.

Korpusi se, dakle, mogu klasificirati prema različitim kriterijima:

- **Jezični varijetet.** Prema ovome kriteriju razlikujemo opće i specijalizirane korpuse pri čemu opći korpusi, kao što su primjerice *HNK* ili *hrWaC*[20], pokrivaju jezik u cijelosti, dok specijalizirani korpusi pokrivaju samo jedan jezični varijetet ili funkcionalni stil (primjerice korpus *ENGRI* (Bogunović i dr., 2021) pokriva publicistički funkcionalni stil). Dok su veći korpusi sastavljeni za istraživanja općenitih jezičnih pojavnosti, specijalizirani su korpusi obično izrađeni kako bi odgovorili na specifična istraživačka pitanja ili pokrivaju određenu domenu. U ovu kategoriju valja ubrojiti i referentne korpuse koji su najbliži „sada već zastarjelu pojmu reprezentativnoga korpusa, a temelje se na nekim relevantnim parametrima oko kojih je postignut lingvistički dogovor te obuhvaćaju i pisani i govoreni jezik, formalne i neformalne njegove registre itd.“ (Bratanić

[19] Riječi *godina, čovjek, dan, vrijeme* najfrekventnije su imenice u *hrWaC*-u.

[20] U *hrvatskome mrežnom korpusu* (*hrWaC*) i *Hrvatskome nacionalnom korpusu* (*HNK*) nisu zastupljeni svi funkcionalni stilovi hrvatskoga jezika te ga se stoga ne može smatrati uravnoteženim i reprezentativnim u smislu u kojemu je to, primjerice, *British National Corpus*.

1998: 174). Ovdje valja spomenuti i tzv. kontrolni korpus (engl. *monitor corpus*) koji se odnosi na korpus „kojem je svrha neprestano ažuriranje referentnoga korpusa kako bi on mogao odražavati jezične mijene, čuvajući pri tom njegovu ravnotežu i sastav“ (Bratanić 1998: 174). Primjer je takva korpusa za engleski jezik *The Bank of English*. U pravilu se takav korpus nadograđuje jednom godišnje, a za hrvatski bi jezik takvo što bilo od velika značaja. Postoje i tzv. oportunistički korpusi (v. također poglavlje 1.3) koji mogu biti jeftina inačica referentnih korpusa, a nastaju za potrebe pojedinih istraživanja kada veličina i reprezentativnost nisu od velika značenja (Teubert, 1995).

- **Medij.** Prema mediju razlikuju se korpusi govornoga jezika (npr. *Hrvatski korpus govornoga jezika*[21]) od korpusa pisanoga jezika (npr. *hrWaC*). Postoje međutim i višemedijski korpusi, prije svega za engleski jezik[22] (Szudarski, 2017).

- **Raspon.** Prema rasponu korpusi mogu biti sinkronijski ili dijakronijski. Korpus je dijakronijski ako je u njemu označeno vrijeme nastanka teksta, odnosno ako omogućuje dijakronijsko istraživanje dane jezične pojave. Međutim, teško je odrediti točnu vremensku granicu kojom tekstovi moraju biti obuhvaćeni jer dijakronijskim korpusom možemo smatrati i korpus koji obuhvaća manja vremenska razdoblja (primjerice korpus u kojemu se proučava nastanak novih riječi u vrijeme pandemije izazvane COVID-om, što može biti svega nekoliko mjeseci ili tjedana), ali i duža vremenska razdoblja (primjerice korpus u kojemu se proučava priljev stranih riječi u hrvatskome u posljednjih dvadeset godina). Primjer korpusa koji je uspostavljen za potrebe jezičnopovijesnih istraživanja hrvatskoga jezika je *Regensburški dijakronijski korpus hrvatskoga jezika – CroDi* (v. tablicu 1).

- **Broj jezika.** S obzirom na broj jezika korpusi mogu biti jednojezični i višejezični. Višejezični korpusi dijele se na usporedne korpuse (engl.

[21] O problemima uzorkovanja specijaliziranih korpusa govornoga i pisanoga jezika odraslih vidi Kuvač Kraljević i dr. (2016).

[22] Usporedbe radi, za engleski jezik razvijeni su alati poput *BNClab* odnosno laboratorij neformalnoga govornog jezika u kojemu je moguće pretraživati pojavnice u korpusu prema spolu govornika, društvenome sloju kojem govornik pripada, regionalnim karakteristikama, i dr. Jedan od primjera multimodalnoga korpusa jest i korpus koji je nastao pri Sveučilištu Nottingham (Knight, 2011, Knight i dr., 2009), koji je sastavljen od internetskih prijenosa (engl. *video streaming*) akademskih razgovora između mentora i studenata, a sadrži anotaciju aspekata govora tijela, što je istraživačima omogućilo da proučavaju odnos između jezika i gesta.

parallel corpora), koji obuhvaćaju tekstove na dva ili više jezika i sastoje se od izvornika i njihovih prijevoda, te usporedive korpuse (engl. *comparable corpora*), odnosno višejezične korpuse sastavljene prema istim parametrima i kriterijima.

Postoje i drugi parametri ili kriteriji prema kojima se može odrediti vrsta korpusa, kao što su opseg, veličina, izvornost tekstova itd. (v. Szudarski, 2017). Prema posljednjemu kriteriju izdvajaju se pedagoški korpusi, odnosno zbirke jezičnoga izričaja, koje su nastale u učioničkome okruženju, a sastoje se primjerice od udžbenika, transkripata iz nastave i sl. Takav korpus može se rabiti kao resurs za procjenu napretka učenika (usp. Nedić 2017: 14). Uz pojam pedagoškoga korpusa veže se i pojam učeničkoga korpusa[23] odnosno elektroničke zbirke tekstova koje su proizveli učenici jezika kojima dotični jezik nije materinski (Mikelić Preradović i dr., 2015). Takvi korpusi pružaju dokaze o učenju jezika te prikazuju načine na koji se jezikom služe neizvorni i izvorni govornici ili govornici s drugim materinskim jezikom (Nedić, 2017).

Uz sažeti prikaz podjele korpusa na vrste, u nastavku se osvrćemo na prednosti i nedostatke mrežnih korpusa.

1.6. MREŽNI KORPUSI

S obzirom na rast interneta i broja mrežnih stranica, pojedini jezikoslovci predlažu uporabu mreže kao korpusa (engl. *Web as Corpus*, Baroni i Bernardini, 2004; Baroni i dr., 2006; Boulton, 2015; Fletcher, 2012; Hundt i dr., 2015; Jakubiček i dr., 2020, i dr.).

Mrežni korpusi su, dakle, velike zbirke (moguće isprekidanoga) teksta s određenoga dijela mreže, s minimalnom količinom metapodataka kao što su izvor i vrijeme preuzimanja (Ljubešić, osobna komunikacija), koji istraživačima na raspolaganje stavljaju veliku količinu jezičnih podataka, a na istraživačima je da procijene korisnost takva korpusa za svoja istraživanja. Najveća su prednost mrežnih korpusa veličina i raznolikost tekstova (usp. također Fletcher,

[23] *Učenički korpus hrvatskoga jezika* (*CroLTeC*) opisan je u tablici 1. O sastavu korpusa i zadacima obrade korpusa v. Mikelić Preradović i dr. (2015).

2012) te pristupačnost i cijena pa ih se često koristi u nedostatku drugih korpusa, kao što je slučaj s hrvatskim jezikom[24].

Velika prednost mrežnih korpusa mogućnost je uporabe velike količine podataka[25] (McEnery i Hardie, 2012), a kao posljedica toga problemi veličine i reprezentativnosti blijede, pa je takav korpus ujedno i kontrolni korpus jer se neprestano nadograđuje ili ažurira. Najčešća metoda prikupljanja takva korpusa je tzv. *crawling*[26] (v. primjerice Ljubešić i Erjavec, 2011; Ljubešić i Klubička, 2014). Jakubiček i dr. (2020) navode nedostatke mrežnih korpusa kao što su strojno prevedeni tekstovi koji ozbiljno mogu utjecati na kvalitetu mrežnoga korpusa jer ne predstavljaju *prirodni jezik*, potom neželjene poruke (engl. *spam*), zatvoreni sadržaj odnosno nedostupnost određenih mrežnih stranica iz političkih ili ekonomskih razloga, dinamičnost sadržaja na mreži što utječe na brzinu povlačenja mrežnih stranica, i dr.

Jedan je od temeljnih nedostataka mrežnoga korpusa, dakle, nemogućnost kontrole kvalitete tekstova koji čine korpus, stoga se tekstovi dobiveni metodom *crawlinga* (v. fusnotu 26) moraju smisleno sortirati da bi se mogli smatrati korpusom. Nadalje, takav korpus sadrži veliku količinu šuma (automatski izbornici, nejezični podaci, tekst niske kvalitete, zatipci, dvostruki podaci itd.), manjak metapodataka, a zbog promjenjive prirode interneta pretrage je teško replicirati, što u pitanje dovodi valjanost istraživanja. No, s druge strane, potonji argument vrijedi i za bilo koji kontrolni korpus, a fluktuacije i promjene u korpusu mogu se sagledati i u pozitivnome svjetlu u smislu da predstavljaju stanje jezika u njegovoj promjenjivosti (Boulton, 2015).

[24] Među južnoslavenskim jezicima slovenski ima najrazvijenije jezične tehnologije. Ovdje ćemo spomenuti referentni kontrolirano izrađeni korpus *Gigafida Corpus* koji predstavlja pisani standardni jezik, sadrži 1,8 milijardi pojavnica i redovito se obnavlja.

[25] Jones i dr. (2007) pokazali su prednosti mrežnih korpusa u istraživanju antonima.

[26] U računalnoj lingvistici rabe se dvije metode prikupljanja tekstova, a to su tzv. *web crawling* i *web scraping*. Prva metoda, tzv. puzanje, odnosi se na preuzimanje (svih) dostupnih *HTML* dokumenata s dijela mreže (.hr), pri čemu je riječ o „oportunom preuzimanju svega na što se naiđe“ (Ljubešić, osobna komunikacija), a potonja, koja je osim po nazivu *web scraping* poznata i pod nazivom *text extraction*, odnosi se na crpljenje glavnoga teksta tj. preuzimanje odlomaka s najviše teksta, što znači da se ne crpe naslovi, metapodaci (autor, vrijeme objave, kategorija…) i sl. Problem je ovakva načina prikupljanja podataka kako identificirati i razlikovati bliske jezike kao što su hrvatski i bosanski, ili hrvatski i srpski. Nakon procesa prikupljanja teksta slijedi uklanjanje (bliskih) duplikata primjerice pomoću alata *onion deduplication tool* (Baroni i dr., 2006); pri čemu se katkad ukloni i dio teksta, što može utjecati na cjelovitost dokumenta. U sljedećoj fazi slijedi obilježavanje teksta (v. 2.).

1.7. RAČUNALNOJEZIKOSLOVNI RESURSI I ALATI ZA HRVATSKI JEZIK

Začetak korpusne lingvistike u Hrvatskoj može se vremenski smjestiti u šezdesete godine 20. stoljeća, dakle u razdoblje koje je paralelno začetku korpusne lingvistike u Velikoj Britaniji i SAD-u (v. 1.1.). Naime, 1967. godine pri Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu nastaje prvi računalni korpus u Hrvatskoj - *Barokni ep Ivana Gundulića Osman*, koji priređuje, frekvencijski obrađuje i konkordancijama popraćuje Željko Bujas (Tadić, 1997). Tadić (1997) navodi da, godinu dana nakon objave korpusa *Brown* (Francis i Kucera, 1967), R. Filipović u sklopu kontrastivnoga projekta *The Yugoslav Serbo-Croatian-English Contrastive Project* (1969), prevodeći pola korpusa *Brown* na hrvatski jezik, stvara prvi računalni usporedni korpus u povijesti svjetske lingvistike koji se rabi za poredbenolingvistička istraživanja (usp. Tadić (1997: 388); Tadić (2009: 219)). Iako je začetak korpusne lingvistike u Hrvatskoj bio impresivan, danas hrvatski zaostaje za mnogim svjetskim i slavenskim jezicima zbog relativno malog broja govornika, slabu zainteresiranost tržišta (Bratanić, 1998) i mali broj istraživača koji se bave korpusnom lingvistikom (Bekavac, 2001). Na početku 21. stoljeća stanje i razvoj korpusne lingvistike u Hrvatskoj bili su daleko iza svjetskih dostignuća, no 15-ak do 20 godina kasnije već postoji značajan napredak, kao što je prikazano u nastavku ovoga poglavlja.

Najveće prekretnice u razvoju korpusne lingvistike na kraju 20. i početkom 21. stoljeća u Hrvatskoj pripisuju se dvama resursima: *Hrvatskome čestotnom rječniku* (Moguš, Bratanić i Tadić, 1999) i *Hrvatskome nacionalnom korpusu* (*HNK*). Prikupljanje *Jednomilijunskoga korpusa hrvatskoga književnog jezika* koji se naziva i *Moguševim korpusom* započelo je 1976. pod vodstvom M. Moguša, a na temelju tog korpusa sastavljen je *Hrvatski čestotni rječnik*. Sastavljanje *Hrvatskoga nacionalnog korpusa* (*HNK*) započeo je M. Tadić 1998. godine, po uzoru na *British National Corpus* (*BNC*) i Češki nacionalni korpus (*CNC*). Smatra se da je *HNK* treći nacionalni korpus u povijesti, a prvi zbog svoje dostupnosti pretraživanja putem interneta (usp. Kolarović i Tadić 2013: 10). Opseg od 101 milijun riječi-pojavnica *HNK* je dosegao 2004. godine u inačici HNKv2.0 i tako postao korpus treće generacije[27] (Tadić, 2009; Kolarović i Tadić, 2013). Velik iskorak učinjen je 2008., od kada je dostupna lematizirana i morfološki označena inačica *HNKv2.5* (Tadić 2009: 221–224)[28].

[27] Korpusima prve generacije smatraju se korpusi koji su nastali u razdoblju od 1967. do 1985., a sadržavali su oko milijun pojavnica. Primjer su takva korpusa korpusi *Brown* i *LOB*. Korpusima druge generacije smatraju se korpusi koji su nastali u razdoblju od 1985. do 1990., a sadržavali su od 10 do 20 milijuna pojavnica. Primjer takva korpusa je korpus *COBUILD*, dok su korpusi treće generacije nastali nakon 1990. god., a sadrže nekoliko stotina milijuna pojavnica. Korpusi su treće generacije, primjerice, *BNC, CNC* i *HNK*.

[28] Usp. također Lalli Paćelat (2014).

HNK je u inačici *v3.0* dosegao granicu od 200 milijuna riječi-pojavnica, te se smatralo da zbog svoje usustavljenosti, uravnoteženosti i opsega predstavlja jedini referentni korpus pisanoga hrvatskog standardnog jezika i ujedno nezaobilaznu građu za sve vrste lingvističkih istraživanja i leksikografskih obradbi suvremenoga hrvatskog jezika (Kolarović i Tadić 2013: 11). Od 2015. godine HNK je instaliran u alat *NoSketch Engine* (*NoSkE*), čime je postignut jedan od osnovnih ciljeva korpusa uopće a to je slobodno, udaljeno, neograničeno pretraživanje jezične građe (usp. Tadić, 2001; također navedeno u Posavec, 2017). Uz *HNK* postoje još dva korpusa hrvatskoga općeg jezika, a to su *Hrvatska jezična riznica*, koja uključuje pisane tekstove od 11. stoljeća do danas, te *Hrvatski mrežni korpus* (*hrWaC,* Ljubešić i Erjavec, 2011) opsega 1,2 milijardi riječi-pojavnica koji predstavlja trenutno najopsežniji korpus hrvatskoga jezika. Korpus *Riznica* s oko 100 milijuna pojavnica obuhvaća temeljna djela hrvatske književnosti (npr. romane, pripovijetke, dramu, poeziju i dr.), publicističke tekstove, znanstvene tekstove, sveučilišne, osnovnoškolske i srednjoškolske udžbenike, književne prijevode, mrežno dostupan tisak, te knjige iz predstandardnoga razdoblja hrvatskoga jezika koje su prilagođene današnjemu hrvatskom jezičnom standardu (usp. također Posavec 2017: 31).

Trenutno najopsežniji korpus hrvatskoga jezika *hrWaC* sastavljen je metodom *crawlinga* (v. fusnotu 26) top domene .hr, kojom je obuhvaćeno 8388 URL-a. Paralelno uz korpus za hrvatski jezik izrađeni su i korpusi za srpski i bosanski jezik (Ljubešić i Klubička, 2014). *HrWaC* je od velike važnosti za razvoj jezičnih tehnologija za hrvatski jezik. Korpus sadrži tekstove standardnoga hrvatskog jezika, kao što su, primjerice, tekstovi povučeni s mrežnih stranica službenih i javnih tijela, ali i tekstove nestandardnoga jezika, primjerice blogove, reklame, rasprave i sl.[29] Prilikom razvoja korpusa *hrWaC* Ljubešić i Erjavec (2011) su proveli usporedbu zastupljenosti tema u hrvatskome i slovenskome korpusu s britanskim mrežnim korpusom *ukWaC* te utvrdili vrlo visok stupanj zastupljenosti sličnih tema u navedenim korpusima (usp. također Posavec, 2017).

Kao što je već navedeno, hrvatski jezik pripada jezicima sa slabije razvijenim jezičnim tehnologijama (Tadić, 2003; Tadić, Brozović-Rončević i Kapetanović, 2012). Jezične tehnologije dijele se na jezične resurse, jezične alate i komercijalne proizvode (Tadić, 2003). Jezični izvori ili resursi predstavljaju jezičnu građu koja je digitalno usustavljenja i služi za pretraživanje. Postoje dva oblika

[29] O prednostima i nedostacima mrežnih korpusa v. 1.6.

jezičnih resursa: korpusi i jezične zbirke, koji daju značajnu količinu jezičnih podataka, te digitalni rječnici, koji se mogu pretraživati mrežno ili izvanmrežno. Jezični su alati, s druge strane, specijalizirani programi, razvijeni na temelju jezičnih izvora, koji omogućuju obradu postojećih jezičnih izvora ili pak stvaranje novih. Vezani su uz jezičnu razinu, pa su tako, primjerice, na fonološkoj razini razvijeni N-grami, na morfološkoj generatori, analizatori, lematizatori, označivači, na sintaktičkoj *parseri*, tzv. *chunkeri*, banke stabala itd., a na semantičkoj *FrameNet* i *WordNet*. Komercijalni su proizvodi rječnici, provjernici (pravopisa, gramatike, stila), sustavi za diktiranje, strojno (potpomognuto) prevođenje (engl. *Machine-aided Translation,* M(A)T) i računalno potpomognuto učenje jezika (engl. *Computer-assisted language learning*, CALL) (Tadić, 2003).

U Hrvatskoj se svega nekoliko ustanova bavi razvojem jezičnih tehnologija. Prije svega valja spomenuti Katedru za algebarsku i računalnu lingvistiku Odsjeka za lingvistiku pri Sveučilištu u Zagrebu, iz koje su proizašli brojni resursi za hrvatski jezik (npr. *Hrvatski nacionalni korpus, Hrvatski morfološki leksikon, Hrvatska ovisnosna banka stabala*[30]), kao i *Hrvatsko društvo za jezične tehnologije*, portal *Jezične tehnologije za hrvatski jezik*, hrvatski *META-SHARE* čvor, *Hrvatski jezik na internetu - JEZIK.Hrvatski,* i dr. Također, Odsjek za informacijske i komunikacijske znanosti poznat i pod nazivom *Natural Language Processing Group* pri Sveučilištu u Zagrebu u novije vrijeme donosi vrijedne resurse kao što su *Stemmer for Croatian, IITerm* i dr., i njihova se internetska stranica najčešće ažurira. Tu je i Fakultet elektrotehnike i računarstva pri Sveučilištu u Zagrebu kao važan centar za razvoj jezičnih tehnologija, posebice Zavod za elektroniku, mikroelektroniku, računalne i inteligentne sustave iz kojega su proizašli alati poput *TermeX*, *Ispravi.Me - Hrvatski akademski spelling checker*, i dr. Značajan je i doprinos Odjela za opće jezikoslovlje pri Institutu za hrvatski jezik i jezikoslovlje s resursima kao što su *Baza frazema hrvatskoga jezika*, *Kolokacijska baza hrvatskoga jezika, Hrvatski mrežni rječnik, Struna - Hrvatsko strukovno nazivlje* i dr. Nadalje, kao vrijedan jezični resurs, iako ne korpus, Posavec (2017) izdvaja i *Portal Leksikografskoga zavoda* koji sadrži 60 652 članaka, čija se baza neprestano nadopunjava. Ovdje valja spomenuti i korpus *ENGRI* (Bogunović i dr., 2021) koji je sastavio riječki tim unutar projekta *Engleske riječi u hrvatskome jeziku: identifikacija, afektivno-semantičko normiranje i ispitivanje kognitivne obrade*

[30] *Resursi - HR4EU.*

bihevioralnim i neuroznanstvenim metodama (HRZZ, 2020 – 2025) pod vodstvom dr. sc. Irene Bogunović.

Od svibnja 2022. godine dostupan je hrvatsko-engleski usporedni mrežni korpus Croatian web corpus *MaCoCu-hr* 1.0 (Bañón, 2022), s 2,3 milijarde pojavnica u 7 milijuna tekstova. Korpus je sastavljen metodom tzv. puzanja (engl. *crawling*, v. fusnotu 26) top level domene .hr u 2021. godini, a koji je za razliku od *hrWaC*-a proširen i na druge domene. Pri tome se znatno pazilo na pročišćavanje i crpljenje teksta kako bi se u konačnici dobio visokokvalitetan mrežni korpus. To je postignuto na način da je uklonjen standardni kod[31] (engl. *boilerplate*), dupli sadržaj[32], i isključeni su jako kratki tekstovi te tekstovi koji nisu iz ciljnoga jezika. Nadalje, uzorci s najvećih 1500 domena ručno su pregledani te su uklonjene domene koje ne zadovoljavaju uvjete kvalitete teksta (kao npr. strojno prevedeni tekstovi). Naposljetku je korpus strojno obilježen (v. 2.) te sadrži cijeli niz metapodataka čime je omogućena pretraga teksta prema kvaliteti te drugim kriterijima[33]. Zbog svega navedenog, ovaj je korpus znimno vrijedan računalnojezikoslovni resurs za potrebe korpusne lingvistike, kao i za uvježbavanje jezičnih modela i općenito za razvoj jezičnih tehnologija za hrvatski jezik.

Od važnih resursa za južnozapadnoslavenske jezike valja spomenuti infrastrukturu *Common Language Resources and Technology Infrastructure* (*CLARIN ERIC*), koja objedinjuje digitalne alate i resurse u cilju omogućavanja dostupnosti jezičnih podataka i alata istraživačima iz različitih znanstvenih disciplina, s naglaskom na humanističke i društvene znanosti, kao i *ReLDI -Regional Linguistic Data Initiative*. Obje infrastrukture paralelno razvijaju jezične resurse i alate za hrvatski, srpski, slovenski i druge južnoslavenske jezike.

Korpusi *hrWaC, Riznica* i *HNK* instalirani su u besplatan računalnojezikoslovni alat *NoSkE*[34] te su dostupni preko *ReLDI servisa* (Samardžić i dr., 2015). Za potrebe rada i studija prikazanih u ovoj knjizi korišten je alat *Sketch Engine* (*SkE*, Kilgarriff i dr., 2004). *Sketch Engine* jedan je od najrasprostranjenijih alata zbog pristupačnosti i mnoštva funkcionalnosti koje nudi korisnicima, koji uz funkcionalnosti koje nudi *NoSkE* nudi još i dodatnu funkcionalnost skica riječi (v. 3), i naposljetku, ne manje važno, pruža dobru podršku korisnicima.

[31] V. *Corpus Tools.*

[32] *V. Onion. Corpus Tools.*

[33] *V. Github CLARINSI/CLASSLA.*

[34] *NoSketch Engine*.

Alat[35] je odabran jer podržava hrvatski jezik i jer predstavlja napredan alat za provođenje jezikoslovnih istraživanja (Kilgarriff i Kosem, 2012). Svi su korpusi na kojima se radilo i koji su prikazani u ovoj knjizi ili već javno dostupni u *SkE*-u ili dostupni na zahtjev, čime se osiguravaju pouzdanost i valjanost provedenih istraživanja.

U *SkE*-u trenutno je dostupno devet korpusa hrvatskoga jezika, od toga su dva korpusa općega jezika (*hrWaC, Riznica*), pet paralelnih korpusa (*OPUS*, *DGT Croatian*, *EUR-Lex Croatian 2/2016*, *EUR-Lex judgements Croatian 12/2016*, *OpenSubtitles 2018 Croatian*), dva korpusa govornoga jezika (*Croatian ParlaMint* i *CHILDES Croatian Corpus*). U *NoSkE*-u dostupno je trinaest korpusa hrvatskoga jezika, dok ih je za slovenski gotovo šezdeset. Usporedbe radi, za engleski jezik postoji stotinjak korpusa u *SkE*-u. Zanimljivo je da je za ostale svjetske jezike kao što su njemački, francuski i španjolski taj broj znatno manji, no pretpostavlja se da i ti jezici imaju jako dobro razvijene jezične tehnologije, no svoje resurse uglavnom dijele na vlastitim platformama (npr. *Digitales Wörterbuch der Deutschen Sprache, DWDS*).

Zaključno, s obzirom na manjak financijskih sredstava i broja istraživača koji se bave korpusnom lingvistikom u Hrvatskoj, hrvatski u razvoju jezičnih tehnologija zaostaje za drugim jezicima. Premda broj korpusa za hrvatski jezik prema podacima iz slike 1 izgleda poražavajuće, stvarno stanje ipak je nešto bolje (v. tablicu 1 s popisom postojećih korpusa hrvatskoga jezika i kategorizacijom prema vrsti, veličini i podacima o autoru/autorima te pristupu korpusu), iako se za sada, nažalost, većina resursa razvija uglavnom individualno i bez velike povezanosti i umreženosti (Ljubešić, osobna komunikacija). No, veliki iskorak u tom smislu učinila je grupa *CLASSLA* (*The CLARIN Knowledge Centre for South Slavic languages*), a za hrvatski jezik posebno Nikola Ljubešić.

[35] Popis više od 265 alata za sastavljanje i analizu korpusa dostupan je na *Corpus Analysis Tools.*

Tablica 1: *Korpusi hrvatskoga jezika* [36] *(abecedni popis, zadnji pristup studeni 2021).*

Red. br.	Naziv korpusa	Skraćeni naziv	Vrsta korpusa	Autor	Veličina	URL
1.	*24sata news article archive 1.0*	/	specijalizirani korpus	Purver, Shekhar, Pranjić, Pollak i Martinc, 2021	657 806 tekstova	https://www.clarin.si/repository/xmlui/handle/11356/1410
2.	*Annotated corpus of Croatian language-related news articles*	*MetaLangNEWS-Hr*	specijalizirani korpus	Bogetić i Batanović, 2020	555 890 pojavnica	https://www.clarin.si/repository/xmlui/handle/11356/1369
3.	*Comparable corpora of South-Slavic Wikipedias*	*CLASSLA-Wikipedia 1.0*	usporedni višejezični korpus	Ljubešić, Marković, Markoska i Erjavec, 2021	486 258 862 pojavnica	https://www.clarin.si/repository/xmlui/handle/11356/1427
4.	*Croatian corpus of DGT*[37]*-Translation Memory*	*DGT*	usporedni višejezični korpus	/	5 123 494 pojavnica	https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fdgt__sh_hr
5.	*Croatian corpus of non-professional written language by typical speakers and speakers with language disorders*	*RAPUT 1.0*	specijalizirani korpus	Kuvač Kraljević, Hržica, Štefanec, Kolografić Belić i Ljubešić, 2021	426 187 pojavnica	https://www.clarin.si/repository/xmlui/handle/11356/1435
6.	*Croatian error-annotated corpus of non-professional written language*		specijalizirani korpus	Štefanec Ljubešić i Kraljević, 2016	oko 500 000 pojavnica	
7.	*Croatian twitter training corpus ReLDI-NormTagNER-hr 2.1*		ručno anotirani korpus tvitova	Ljubešić, Erjavec, Batanović, Miličević i Samardžić, 2019	89 104 pojavnica	https://www.clarin.si/repository/xmlui/handle/11356/1241
8.	*Croatian-English parallel corpus*	*hrenWaC 2.0*	usporedni korpus	Ljubešić, Espià-Gomis, Ortiz Rojas, Klubička i Toral, 2016	55 083 246 riječi	https://www.clarin.si/repository/xmlui/handle/11356/1058

[36] Ostali resursi za hrvatski jezik, poput *Kontekst.io,* koji su dostupni preko servisa *CLARIN*.

[37] Sintaktički anotirana inačica korpusa dostupna je i na repozitoriju *CLARIN*.

Red. br.	Naziv korpusa	Skraćeni naziv	Vrsta korpusa	Autor	Veličina	URL
9.	*Croatian-English parallel corpus*	*MaCoCu-hr-en 1.0*	usporedni korpus	Bañón, i dr., 2022	134 850 790 riječi	http://hdl.handle.net/11356/1522
10.	*EUR-Lex Croatian 2/2016*	*Eur_Lex Cro*	usporedni višejezični korpus	/	156 309 317 pojavnica	https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Feurlex_hrv
11.	*EUR-Lex Judgements Croatian*	*EUR-Lex Judgements Cro*	usporedni višejezični korpus	/	7 416 811 pojavnica	https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fjudgments_eurlex_hrv
12.	*Hrvatska jezična riznica*	*Riznica 1.0, HJR*	korpus standardnoga hrvatskog jezika	Brozović i dr., 2018	101 782 863 pojavnica	http://riznica.ihjj.hr/index.hr.html
13.	*Hrvatski korpus dječjega jezika*	*HKD, CHILDES Croatian*	specijalizirani korpus, korpus govornoga jezika	Kovačević, 2002	389 674 pojavnica	http://childes.psy.cmu.edu/data/Slavic
14.	*Hrvatski korpus govornoga jezika*	*HrAL*	korpus govornoga jezika	Kraljević i Hržica, 2017	250 000 pojavnica	https://ca.talkbank.org/access/Croatian.html
15.	*Hrvatski mrežni korpus*	*hrWaC*[38]	korpus općeg jezika	Ljubešić i Erjavec, 2011	1 397 757 548 pojavnica	https://www.sketchengine.eu/hrwac-croatian-corpus/?gclid=Cj0KCQiAsqOMBhDFARIsAFBTN3dXWXfTc4yXM0gS_rMg9B2lAgVdI2Q5RoaIdT015eWjhXUVVV6hZ4gaApiPEALw_wcB
16.	*Hrvatski nacionalni korpus*	*HNK_v30*	korpus općega standardnog jezika	/	2 559 160 riječi	http://filip.ffzg.hr/cgi-bin/run.cgi/first_form http://hnk.ffzg.hr/

[38] *hrWaC* je u *SkE*-u dostupan u dvije inačice, RELDI i *RF Tagger*.

Red. br.	Naziv korpusa	Skraćeni naziv	Vrsta korpusa	Autor	Veličina	URL
17.	*Hrvatski učenički korpus*	*CroLTeC*	specijalizirani korpus	Mikelić Preradović i dr., 2015	1 054 287 pojavnica	https://www.bib.irb.hr/795386
18.	*Korpus hrvatskih internetskih portala (2014-2018)*	*ENGRI*	specijalizirani korpus	Bogunović, Kučić, Ljubešić i Erjavec, 2021	694 799 268 pojavnica	https://www.clarin.si/repository/xmlui/handle/11356/1416
19.	*Korpus medijskog metajezika*	*MetaLangCORP-NEWS-hr*	specijalizirani korpus	Bogetić i dr., 2021	535 455 pojavnica	
20.	*Multilingual comparable corpora of parliamentary debates*	*ParlaMint 2.1*	specijalizirani korpus, usporedivi korpus, korpus govornoga jezika	Erjavec i dr., 2021	494 949 904 riječi	http://hdl.handle.net/11356/1432
21.	*Offensive language dataset of Croatian, English and Slovenian comments*	*FRENK 1.0*	specijalizirani korpus	/	32 795 tekstova	https://www.clarin.si/repository/xmlui/handle/11356/1433
22.	*Open Source Parallel Corpus, Croatian*	*OPUS 2 Croatian*	usporedni višejezični korpus	/	156 942 211 pojavnica	https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fopus2_hr
23.	*Regenburški dijakronijski korpus hrvatskoga jezika*	*CroDi*	dijakronijski korpus	Hansack, Hansen, Horvat i Perić Gavrančić, 2016	820 317 pojavnica	http://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/rund-ums-institut/regensburger-korpora/index.htm http://huslaw01.german.hu-berlin.de/welcome/default/index

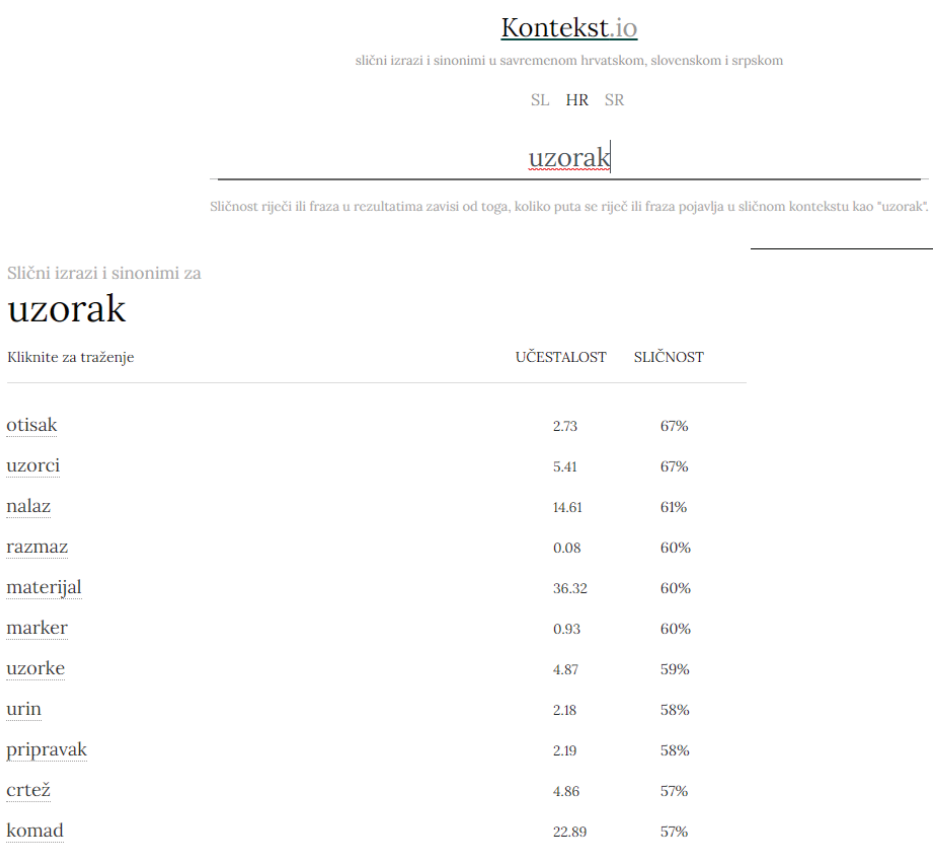
Red. br.	Naziv korpusa	Skraćeni naziv	Vrsta korpusa	Autor	Veličina	URL
24.	Semantic hyper-graph corpus	SemCRO 1.0	korpus za uvježbavanje	Vasić i dr. 2020	184 rečenice	https://www.clarin.si/repository/xmlui/handle/11356/1377
25.	Tourism English-Croatian Parallel Corpus 2.0	/	specijalizirani usporedni korpus	Toral i dr., 2016	139 938 unosa	https://www.clarin.si/repository/xmlui/handle/11356/1049
26.	Training corpus hr500k[39]	hr500k 1.0	referentni označeni korpus	Ljubešić, Agić, Klubička, Batanović i Erjavec, 2018	500 000 pojavnica	http://hdl.handle.net/11356/1183 https://www.clarin.si/noske/run.cgi/corp_info?corpname=hr-500k&struct_attr_stats=1
27.	Twitter sentiment for 15 European languages	/	korpus za uvježbavanje	Mozetič, Grčar i Smailović, 2020	1 643 735 jedinica	http://hdl.handle.net/11356/1054

[39] *hr500k* referentni je označeni korpus hrvatskoga jezika sastavljen od 900 dokumenata podijeljenih u 24 794 rečenice, ili 506 457 pojavnica. Predstavlja proširenje prethodnih označenih korpusa hrvatskoga jezika, poput *SETimes.HR* i *SETimes.HR+*. Korpus je ručno segmentiran na pojavnice, rečenice i dokumente, lematiziran je, sadrži morfosintaktičke oznake, te je označen na razini sintakse, semantičkih uloga i imenovanih entiteta. Semantičke uloge su označene u najstarijem dijelu korpusa, odnosno u prva 163 dokumenta odnosno 83 630 pojavnica, koji potječu iz prvobitnoga korpusa *SETimes.HR*. Inačica korpusa *hr500k* može se preuzeti s repozitorija *CLARIN*. Korpusu se također može pristupiti preko *NoSkE-a* i *Kontekst.io*. Postupak izrade korpusa opisan je u Ljubešić i dr. (2018).

1.8. PRETRAŽIVANJE KORPUSA

U ovome potpoglavlju opisujemo načine pretraživanja korpusa te pojašnjavamo razliku između sučelja za pretraživanje i računalnojezikoslovnih alata kojima se može pristupiti korpusu i pretraživati ga na način koji je uobičajen za jezikoslovna istraživanja (npr. prema obliku riječi, lemi i sl.). Sučelja za pretraživanje ili pretraživači, također poznati i pod nazivom tražilice, kao što je *Kontekst.io* (slika 1), omogućuju interakciju između korisnika i stroja, a služe za unos podataka ili naredbi koje se potom prikazuju korisniku (slika 2). U sučelju se stroju da naredba za pretragu na način da se upiše tražena riječ ili izraz (u danome primjeru riječ *uzorak*) za koju program prikaže tražene podatke (tj. sinonime tražene riječi ili izraza na temelju skica riječi dobivenih iz *hrWaC*-a, v. 3).

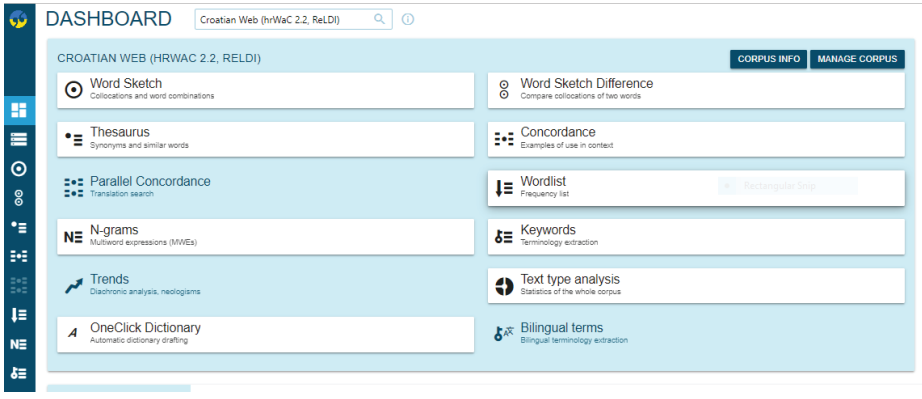
Slika 1: *Prikaz sučelja* Kontekst.io



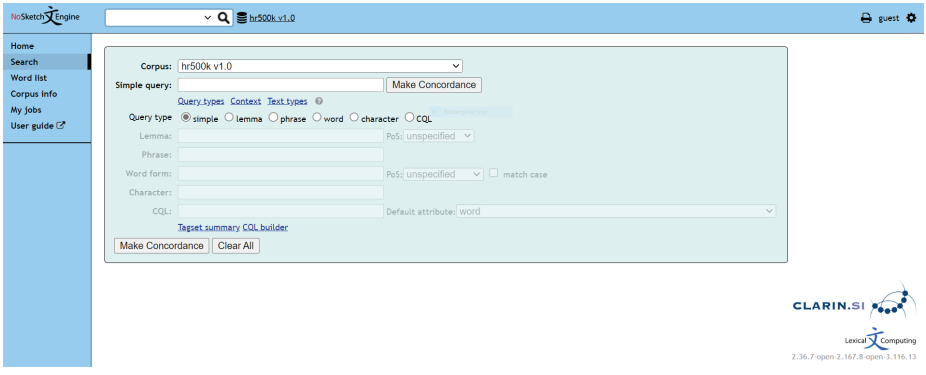
Slika 2*: *Prikaz rezultata za upit uzorak u sučelju Kontekst.io

Dok korisnik sučelju za pretraživanje u pravilu daje jednostavnu naredbu ili upit te jednim klikom dobiva tražene podatke, pristup pojedinome korpusu moguć je jedino pomoću računalnojezikoslovnih alata kao što su *Sketch Engine* (slika 3), *NoSketch Engine* (slika 4), *#LancsBox* (slika 5) i dr. Za razliku od sučelja za pretraživanje, računalnojezikoslovni alati nude znatno više funkcionalnosti i mogućnosti davanja naredbi i dohvata podataka. Računalnojezikoslovni alat *Sketch Engine* (Kilgarriff i dr., 2004) jedan je od najkorištenijih alata zbog činjenice da podržava više od 90 jezika te sadrži više od 500 resursa, odnosno korpusa koji se mogu pretraživati. Korpusi koji su dostupni preko alata *Sketch Engine* u pravilu su veliki, u cilju pokrivanja svih potreba korisnika korpusa. Ostali alati koji se rabe u računalnojezikoslovnim istraživanjima

su: *#LancsBox* (Březina i dr., 2020); *AntConc* (Anthony, 2019a); *WordSmith Tools*; *CQPweb* i dr. Većina alata razvijena je za engleski jezik, dok se alati za jezike koji imaju manji broj govornika, kao što je hrvatski, puno sporije razvijaju (v. 1.7). Za hrvatski jezik dostupna su dva alata, a to su komercijalni alat *Sketch Engine* (*SkE*)[40] te alat u slobodnome pristupu, *NoSketch Engine* (*NoSkE*)[41].



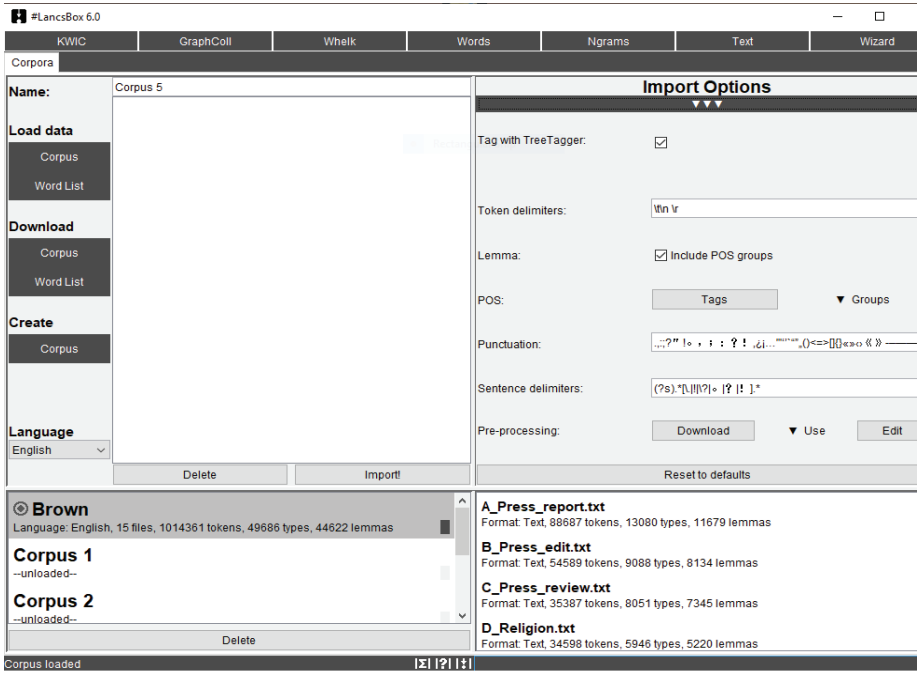
Slika 3: *Prikaz sučelja za pretraživanje u računalnojezikoslovnom alatu* Sketch Engine



Slika 4: *Prikaz sučelja za pretraživanje u računalnojezikoslovnom alatu* NoSketch Engine

[40] *Sketch Engine.*

[41] *NoSketch Engine.*



Slika 5: *Prikaz sučelja za pretraživanje u računalnojezikoslovnom alatu* #LancsBox

Kao što je navedeno u poglavlju 1.4 korpusi su danas postali višemilijunski, odnosno sadrže velike količine podataka, a navedeni alati osmišljeni su i izrađeni upravo da bi se olakšala pretraga korpusa i dohvat podataka iz korpusa. Pretraživanje korpusa bez računalnojezikoslovnoga alata ne bi bilo moguće, a jedna je od najčešćih funkcionalnosti koje alati nude „mogućnost pretraživanja korpusa po upitu, nakon čega program daje najčešće informacije o broju pojavnica u korpusu i njihove primjere u tekstovima u kojima se pojavljuju, a nazivaju se konkordancijski nizovi“ (Posavec 2017: 37, usp. također Nesselhauf, 2005). Ostale funkcionalnosti alata jesu sortiranje, sužavanje rezultata pretrage ovisno o tome što se pretražuje, način preuzimanja i pohrane podataka i dr.

Računalnojezikoslovni alati mogu se svrstati u četiri kategorije (v. tipologiju prema Kilgarriff i Kosem, 2012):

Prva se kategorija odnosi na način na koji se pristupa alatima, tj. je li programski paket potrebno instalirati na računalu ili je mrežno dostupan. Primjer mrežno dostupnih alata jesu *SkE* i *NoSkE*. Alati koji zahtijevaju instalaciju na

računalu su, primjerice, *#LancsBox* (Březina i dr., 2020)*, AntConc* (Anthony, 2019a)*, WordSmith Tools* (Scott, 2008), *MonoConc* (Barlow, 1999), *ParaConc* (Athelstan, 2010).

U drugu se kategoriju svrstavaju korpusno zavisni i korpusno nezavisni alati, pri čemu se prvi koriste isključivo za točno određeni korpus koji je u pravilu nastao unutar projekta određene institucije. Najpoznatiji primjer takva korpusa jest *BNCWeb* tj. korisničko sučelje za pretraživanje *Britanskoga nacionalnog korpusa*. Za razliku od zavisnih alata, nezavisni alati korisniku omogućuju da učita, pretražuje i analizira bilo koji korpus, tj. vlastiti korpus (uz ograničenja za dani jezik), a to su primjerice *SkE* (Kilgarriff i dr. 2004), *NoSkE, #LancsBox* (Březina i dr., 2020)*, AntConc* (Anthony, 2019a)*, WordSmith Tools* (Scott, 2008), *MonoConc* (Barlow, 1999), *ParaConc* (Athelstan, 2010).

U treću kategoriju autori svrstavaju pripremljene korpuse i mrežne korpuse. Pripremljeni korpusi odgovaraju pojmu tradicionalnoga korpusa koji je sastavljen za potrebe istraživanja nekoga jezika, dok se mreža (engl. *web*) koja sadrži velike količine tekstova na različitim jezicima također može smatrati vrstom korpusa (usp. Kilgarriff i Grefenstette, 2003, v. također 1.6.). Pomislili bismo u tom slučaju da se i tražilice kao što su *Google* ili *Bing* također mogu smatrati računalnojezikoslovnim alatima, ali to nije točno jer nisu primarno osmišljene da omogućuju jezikoslovna istraživanja.

U četvrtoj su kategoriji alati svrstani prema složenosti. Jednostavni alati sadrže funkcije poput konkordancija i frekvencija (primjerice *MonoConc,* Barlow, 1999), dok složeniji alati sadrže kolokacije, tezauruse, skice riječi, mogućnost pretraživanja pomoću *CQL*-a i sl. (primjerice *SkE* i *KorpusDK*).

Naposljetku, zadnja kategorija odnosi se na korisnike korpusa, a korisnici su svrstani u tri skupine: 1. leksikografi, 2. istraživači jezika, 3.studenti, nastavnici i učenici.

Pretraživanje korpusa u alatima *SkE-u* i *NoSkE*-u vrši se na način da korisnik preko sučelja (v. slike 3 i 4) postavi upit. Vrste upita (engl. *query type*) su:

1. jednostavan upit (engl. *simple*)
2. lema (engl. *lemma*)
3. izraz ili fraza (engl. *phrase*)
4. oblik riječi (engl. *word*)
5. znak (engl. *character*)
6. CQL.

Prvi je upit, kako sam naziv i upućuje, najjednostavniji, a prilikom ove pretrage korisnik unosi a sustav prikazuje traženu riječ ili riječi čija lema (natuknički oblik, v. 3.) odgovara traženoj riječi. Najčešći su format prikaza rezultata u korpusu ključne riječi u kontekstu (engl. *key word in context*, KWIC), također nazvan konkordancijskim nizom, pri čemu se traženi pojam prikazuje u tekstnom okruženju, tj. uz pojavnicu/pojavnice se prikazuju lijevi i desni ko-tekst (slika 6), no rezultati mogu biti prikazani i kao rečenice. Format KWIC istraživačima omogućuje pregled velike količine primjera. Korisnik može potom odabrati načine i količinu podataka koje želi preuzeti.

Slika 6: *Konkordancijski niz dobiven jednostavnim upitom tražene riječi* uzorak

Pretragom lema dobivaju se pojavnice u svim padežnim oblicima. U slučaju riječi *uzorak* dobivamo isti rezultat kao i jednostavnim upitom jer riječ *uzorak* odgovara samo jednoj lemi, tj. imenici *uzorak*. Međutim, pretraga preko jednostavnoga upita neće polučiti isti rezultat za riječ *riba* jer su toj riječi dodijeljene dvije leme: imenica *riba* i glagol *ribati*. Prilikom pretrage lema možemo podesiti opciju i vrste riječi, npr. lema + imenica te time suziti pretragu za one riječi

kojima je dodijeljeno više lema. Ove pretrage nisu osjetljive na mala i velika slova, a prilikom pretrage mogu se koristiti i regularni izrazi poput zvjezdice, upitnika, i dr. (v. tablicu 2).

Opcija *izraz* nudi traženje višerječnih naziva, a rezultati su prikazani u onom obliku u kojem je izraz unesen odnosno upit postavljen, npr. *uzorak vode* prikazat će samo nominativni oblik, no ne i ostale oblike (*uzorci vode, uzorcima vode* i sl.) koji se pak mogu pretraživati prema određenom obliku riječi ili pomoću CQL-a.

Pretraga oblika riječi omogućuje pretraživanje određene riječi u zadanome morfosintaktičkom obliku, bez njene lematizacije, a ova pretraga daje rezultate za točno određeni oblik riječi, npr. *uzorcima* itd.

Upit znak obuhvaća pretragu određenoga niza znakova te pronalazi pojavnice koje sadrže specifičan znak ili niz znakova. Npr. pretraga niza znakova *uz* pronalazi primjere kao što su ***uz***, *pre**uz**eti*, ***uz**rast*, *pod**uz**eće*, *iz**uz**etan*, ***uz**orak* itd.

Naposljetku, najsloženiji je, ali i najpouzdaniji način pretrage korpusa pretraživanje *CQL*-om (v. 3.) pri čemu se specificira atribut koji se pretražuje (npr. riječi (engl. *word*), lema (engl. *lemma*), oznaka vrste riječi (engl. *tag*), lema i vrsta riječi (engl. *lempos*)), a za sve pretrage moguće je odabrati i opciju *lowercase* čime pretraga postaje osjetljiva na mala i velika slova, pa je tako moguće pretraživati zasebno riječi ili leme *most, Most, MOST*. *CQL* se sastoji od regularnih izraza koji se „koriste nad atributnim izrazima i/ili strukturama“ (Posavec 2017: 48). Najvažniji regularni izrazi[42] za potrebe računalnojezikoslovnih istraživanja s primjerima za hrvatski jezik prikazani su u tablici 2.

[42] Vidi također: *Regular Expressions* | *Sketch Engine.*

Tablica 2: *Regularni izrazi*

Regularni izraz	**Značenje**	**Primjer**	
.	zamjenjuje bilo koji znak samo jednom	l.k bo.	*lak, lik, luk* itd. *bod, bog, boj, bor, bos* itd.
*	označava pojavljivanje prethodnoga znaka 0,1 ili bilo koji veći broj puta	bl*og ko.* *ost	*bog, blog* riječ koja počinje slovom k (npr. *koji, količina, kuća* itd.) riječ koja završava na -ost (npr. *dostupnost, izvrsnost, javnost sigurnost* itd.)
?	označava pojavljivanje prethodnoga znaka 0 ili 1 puta	bl?og	*bog, blog*
+	označava 1 ili više ponavljanja prethodnoga znaka	do+	*do* (prijedlog), *doo* (npr. *dioničko društvo s vlastitom odgovornošću*)
|	ili	bos|kos	*bos* *kos*
( )	grupiranje	(za)?pisati	*pisati, zapisati* Valja spomenuti da se zagrada ne može koristiti sama, već se vrijednosti u zagradi grupiraju kako bi se potom koristila neka druga funkcija. Iz tog razloga je u primjeru naveden upitnik u pretrazi. U navedenome primjeru *(za)pisati* grupira se vrijednost *(za)*, ali ako bismo postavili upit na ovaj način, tj. bez upitnika, dobili bismo zapisati. Grupiranje se u ovome slučaju vrši s upitnikom koji označava da se prethodni znak ne pojavljuje ili pojavljuje jedanput.

[]	raspon svaku pojavnicu potrebno je kod pretrage CQL-om staviti u uglate zagrade	[abcdefghijklmnopqrstuvwxyz] [ABCDEFGHIJKLMNOPQRSTUVWXYZ] [0123456789] [abcde] [lemma="zauzeti"]	[a-z] [A-Z] [0-9] [a-e] *zauzeti, zauzeo je, zauzeli su* itd.
{}	označava ponavljanje znaka ispred zagrade određeni broj puta	a{3} a{3,5} [a-z]{4,} [lemma="zauzeti"] [ ] {2,4}[lemma="stav"]	*aaa* *aaa, aaaa, aaaaa* riječi koje imaju 4 ili više znakova Ono što slijedi iza zareza označava početni i završni broj znakova. U slučaju da iza zareza ne slijedi ništa (npr. [a-z]{4,}), tada je konačni broj znakova jednak broju znakova najduže riječi. primjeri kolokacije *zauzeti stav* s dvije do četiri riječi između npr. *zauzeti jasan i nedvosmislen stav; zauzeo je oštar stav; zauzeli su pravilan stav*
=	oznaka za utvrđivanje vrijednosti atributa	[tag="Np.*"]	vlastite imenice
^	znak koji se izostavlja (potrebno je staviti taj znak u zagradu)	l[^u]k	*lik, lak* (ali ne *luk*)

\	isključivanje elemenata	. \	.(točka kao interpunkcijski znak, a ne točka kao regularni izraz koji zamjenjuje bilo koji znak samo jednom) Omogućuje pretragu znakova koji su dijelom regularnoga izraza kao da to nisu.
(?i)	omogućuje pretragu znakova s velikim i malim slovima	[word="(?i)most"]	*Most, most, MOST*
“ “	oznaka kojom se definira vrijednost	[tag="N."]	sve riječi koje su u korpusu označene kao imenice
&	spaja dva uvjeta	[word="riba"& tag="N.*"]	Pronalazi riječ *riba* koja je imenica, a ne pronalazi primjerice 3. lice glagola *ribati (On riba pod.)*
!	negacija	[word="kapa" & !tag="N.*"]	*Voda kapa u stakleni vrč.*
<>	omogućuje pretragu struktura	<s> [word="(?i)most"]	Pronalazi riječ *most* na početku rečenice. Npr. *Most je raspona 21 metar.*

U tablici 2 navedeni su osnovni regularni izrazi. Iako se njih može pronaći na web stranici *SkE-a* ovdje su navedeni neki primjeri za hrvatski jezik. Složeniji upiti pomoću *CQL-a* prikazani su u poglavljima 5, 6 i 7.

Korpus mora biti strojno obilježen, odnosno označen (v. 2) da bi ga se moglo pretraživati *CQL*-om. Pri tome se pojam koji pretražujemo navodi unutar navodnih znakova (" "), a pojam ili oblik koji se pretražuje je vrijednost atributa. S obzirom na to da svaki atribut ima svoje značenje, odabir pojedinoga atributa dat će i drugačiji rezultat. Pretraga se postavlja na sljedeći način: [word="most"], [lemma="most"] [tag="Ncm.*"] itd. Primjerice, pretraga u obliku [lemma="zauzeti"] [ ] {2,4}[lemma="stav"] računalu daje upute da u korpusu pronađe primjere kolokacije *zauzeti stav* s dvije do četiri riječi između, a rezultat koji ćemo moći iščitati su sintagme poput: *zauzeti jasan i nedvosmislen stav; zauzeo je oštar stav; zauzeli su pravilan stav* itd. Za upite *CQL*-om potrebno je poznavati jezične specifikacije tj. morfosintaktičke oznake za jezik

koji se pretražuje, a one su dostupne pod opcijom Informacije o korpusu – Opće informacije – Skup oznaka (engl. *Corpus Info – General info – Tagset*). Skup oznaka za hrvatski jezik dostupan je i preko *MULTEXT-East Croatian part of speech tagset*.

Često se čini da prilikom pretrage korpusa dobijemo različite brojeve za istu pretragu, primjerice ako uspoređujemo rezultate dobivene preko *konkordancija* i *lista riječi*. Najčešći razlog tomu jest da pretrage nisu identične, a ako jesu, moguće je da se rezultati pretrage razlikuju ovisno o tome jesmo li pretraživali po malim/velikim slovima, lemi ili riječi ili vrsti riječi (primjerice riječ *break* u engleskome može biti i glagol i imenica pa ako pretražujemo prema riječi (*word*), dobit ćemo veći broj pojavnica, a pretraga prema lemi dat će znatno manji broj). Nadalje, engleski član *the* obično se javlja na početku rečenice u engleskome, ali se uglavnom lematizira malim slovima, pa će pretraga [word="the"] dati manje rezultata od pretrage [lemma="the"]. Da bismo doskočili tom problemu, najbolje je koristiti opciju koja podupire pretragu po malim slovima (v. 1.8.) [lemma_lc="the"] pod uvjetom da je korpusu dodan atribut lemma_lc. Isto tako, želimo li dobiti što više rezultata za riječ *and,* bolje je postaviti pretragu na [lc="and" | lemma_lc="and"]. Jednostavne pretrage su korisnicima jednostavne, ali računalu ne, stoga je za optimalne rezultate bolje postavljati upite pomoću *CQL*-a. Valja napomenuti da se rezultat može dobiti korištenjem različitih regularnih izraza (usp. npr. [word="kapa" & !tag="N.*"] i [word="kapa" & tag="V.*"]), iako valja imati na umu da ova dva upita neće dati identičan rezultat.

CQL je moćan alat koji korisnicima pruža mogućnost raznolike i sveobuhvatne pretrage teksta (Posavec, 2017), ali zahtijeva posjedovanje informatičke pismenosti, a također i uloženo vrijeme koje je potrebno da bi korisnik ovladao ovim načinom pretraživanja korpusa.

Rezultate koje dobijemo pretragom korpusa potom promatramo i interpretiramo, pri čemu je bitno da kao korisnici korpusa kritički i kvalitativno promatramo podatke dobivene iz korpusa (usp. također Březina 2018: 19) jer korpusna lingvistika nije samo kvantitativna metoda, i ona bi bez kvalitativne analize bila nepotpuna (Teubert i Čermáková 2007: 124-5). Stoga je ove dvije metode potrebno kombinirati.

Da bismo mogli postavljati upite i dohvaćati podatke iz korpusa kao što je opisano u ovome dijelu knjige, korpus mora biti strojno obilježen, što je tema idućega poglavlja.

2. STROJNO OBILJEŽAVANJE KORPUSA

Nakon što smo u prvome poglavlju dali prikaz korpusne lingvistike kao metodologije te opisali jezične tehnologije za hrvatski jezik, u ovome poglavlju bavimo se problematikom strojnoga obilježavanja korpusa.

Neobilježeni korpus (engl. *raw corpus*) odnosno zbirka tekstova kojoj nisu dodijeljena jezična obilježja nema veliku vrijednost jer nije strojno čitljiv (v. 2., usp. također McEnery i Wilson, 2001) i ne omogućuje složene jezične pretrage. Stoga je obilježavanje jedan od ključnih elemenata korpusa. Obilježavanje računalu omogućuje da čita tekst, što znači da se riječima dodjeljuje niz informacija koje su eksplicitne i koje računalu omogućuju da u konačnici riječi prikaže na način na koji ih mi razumijemo.

Obilježavanje (engl. *annotation, mark-up*) se odnosi na pridodavanje dodatnih eksplicitnih informacija tekstu za računalnu obradu. Obilježavanje može biti tekstno (engl. *textual mark-up, metadata*) ili lingvističko, za što se također rabi naziv lingvističko anotiranje (engl. *linguistic annotation*). Za potonje u literaturi nalazimo i naziv označavanje (usp. Bekavac, 2001)[43] jer se tekstu dodaju metajezične oznake, za razliku od obilježavanja kod kojeg se dodaju tekstna ili kontekstualna obilježja.

Nadalje, naići ćemo u literaturi i na naziv *tegiranje*; prema engleskoj riječi *tagging*, što se odnosi na proces pridruživanja oznaka (engl. *tags*) iz skupa ili popisa oznaka dijelovima teksta (npr. pojavnicama, rečenicama i sl.) odnosno delimitiranim jezičnim jedinicama (usp. Bekavac i Tadić, 2003; Ljubešić i Klubička, 2014; McEnery i Hardie, 2012; i dr.).

Za razliku od atributa i njegova značenja u sintaksi, atribut u korpusnoj lingvistici pretstavlja niz pozicijskih podataka koji se dodaju svakoj pojavnici u korpusu, npr. kojoj vrsti riječi pojavnica pripada, pa tako primjerice imenica ima pet atributa u hrvatskome – opća ili vlastita, rod, broj, padež, živo ili neživo.[44] Osim pozicijskih atributa, postoje i strukturni atributi ili strukture, tj. podaci o strukturama u korpusu ili metapodaci (engl. *metadata*). Strukture čine sastavni dio korpusa, što znači da korpus može biti (i preporučljivo je da bude) podijeljen na manje dijelove kao što su:

[43] Više o obilježavanju i označavanju v. Bekavac (2001).

[44] *MULTEXT-East Croatian part-of-speech tagset (version 5)*.

- <d> - dokument
- <p> - odlomak
- <s> - rečenica
- <g> - spojeni oblici (engl. *glue*)

Primjerice u *SkE*-u struktura <doc> se automatski ubacuje na početku ili na kraju datoteke, mrežne stranice ili dokumenta. Odlomke, tj. <p>, programi teže detektiraju automatski, a uglavnom se ubacuju za sadržaj koji je prikupljen s mreže i odgovara *html* oznaci <p>. Korisnik korpusa ovu strukturu može i ručno dodati tekstu. Struktura rečenice u *SkE*-u automatski se prepoznaje i dodaje korpusu.

Segmentacija rečenica (engl. *sentence segmentation, sentence boundary disambiguation*) „obavlja (se) ubacivanjem jedinstvenih nizova pismena, tj. graničnih oznaka na početak, odnosno na završetak rečenica u tekstu“ (Bekavac 2002: 175).[45] Postoji i posebna oznaka <g> (*glue*) koja prikazuje pojavnice kako ih obično vidimo u pisanome tekstu, npr. riječ *don't* sastoji se od dvije pojavnice *don* i *'t*, iako se korisnicima prikazuje kao jedna.

Prilikom obilježavanja oznake koje se pridodaju tekstu imaju određen redoslijed koji nije striktan i razlikuje se od jezika do jezika, a najčešće se polazi od tokenizacije (tj. opojavničenja), zatim se tekst segmentira na rečenice, nakon čega slijede lematizacija, odnosno dodjela leme svakoj riječi, označavanje vrste riječi (engl. *POS tagging*), morfosintaktičko označavanje (engl. *MSD tagging*), sintaktičko plitko ili dubinsko parsanje, prepoznavanje naziva (engl. *named-entity recognition*, NER (Bekavac 2002: 175), semantičko označavanje (npr. Raysonov[46] (n.d.) programski sustav *Wmatrix Corpus Analysis and Comparison Tool* riječima, uz oznaku vrste riječi, automatski dodjeljuje i semantičku oznaku, čime je omogućena pretraga riječi prema semantičkim poljima (engl. *semantic field*), diskursno označavanje itd. (Garside i dr., 2020)).

Tokenizacija je određivanje pojavnica, ili svega onoga što, pojednostavljeno rečeno, najčešće odgovara onomu što se nalazi između dvije bjeline (Bekavac 2002: 175). No, iako na prvi pogled jednostavan, proces tokenizacije nije uvijek takav. Primjerice, riječ *nemo'š* vidimo kao jednu riječ, iako se ona zapravo sastoji od dviju pojavnica. Nadalje, višerječni nazivi (npr.

[45] Iako se na prvi pogled ne čini da je tako, postupak tokenizacije također je složen jer su interpunkcijske oznake često višeznačne. Točka, primjerice, može stajati na kraju rečenice, uz redni broj, kraticu i sl. (Bekavac 2002: 175).

[46] *Wmatrix Corpus Analysis and Comparison Tool*.

računalno-jezikoslovni alati, računalno jezikoslovni alati i *računalnojezikoslovni alati*) također su problematični, jer, iako predstavljaju jednu leksičku i semantičku cjelinu, pojavljuju se u više inačica te ih računalo u pravilu tretira kao odvojene cjeline, što može utjecati na rezultate dobivene korpusnom pretragom, posebice frekvencije i N-grame (v. 3.).

U postupku lematizacije[47] alat koji se naziva lematizator pojavnice svodi na njihov natuknički oblik. Primjerice glagolski oblici *dođe, došao, dođete* svode se na lemu *doći*, a glagolski oblici *dolaziše, dolazi, dolaze* na glagol *dolaziti*. Na slici 7 naveden je primjer jezične anotacije u kojemu su, uz ključnu riječ u kontekstu (engl. *keyword in context*), prikazane i leme kojima su riječi dodijeljene.

Slika 7: *Prikaz jezične anotacije u korpusu* Riznica

Osim lema, na slici 1 prikazane su i morfosintaktičke oznake za svaku ključnu riječ. Kod označavanja POS i MSD riječima se dodjeljuju vrsta riječi (*kuća*_N, *dolaziti*_G) i morfosintaktički opis (*kuća – Ncfsn*[48], *dolaziti Vmr3s*).[49]. POS označivači[50] mogu biti vrlo precizni jer se iz konteksta u pravilu lako može odrediti gramatička kategorija riječi (Bekavac 2002: 178).

[47] Lematizacija je iznimno složen i važan postupak, posebice za jezike bogate morfologije kao što je hrvatski.

[48] Oznaka *Ncfsn* označava opću (engl. *common*) imenicu (engl. *noun*) ženskoga roda (engl. *feminine*) jedinine (engl. *singular*) u nominativu (engl. *nominative*). Oznaka *VMr3s* označava glagol (engl. *verb*) koji je glavni (engl. *main*), u sadašnjem vremenu (engl. *present*) i 3. licu jednine (engl. *3 singular*). Važno je imati na umu da su navedeni atributi pozicijski te je njihov redoslijed nužno pratiti prilikom složenije pretrage korpusa (usp. CQL, poglavlje 2.2).

[49] O oblikovanju korjenovatelja za hrvatski v. Pandžić (2015).

[50] Analizu uspješnosti morfosintaktičkoga označavanja za hrvatski proveli su Agić i Tadić (2006); Agić, Dovedan i Tadić (2008); Agić, Tadić i Dovedan (2009).

Parseri, jednostavno rečeno, riječi svrstavaju u sintaktičke skupine ili sintagme (npr. imenska, glagolska, pridjevska, i sl.), dok semantički parseri riječi svrstavaju u semantičke kategorije.

Obilježavanje, iako se vrši automatski, ovisi o gramatičkome modelu koji je prihvaćen, pa se tako primjerice participi često označavaju kao pridjevi, kao što je slučaj s riječi *izlazeći*, pri čemu u rečenici *Moli milostinju od sviju izlazećih* particip stoji umjesto odnosne surečenice, te nema ulogu pridjeva već skraćene nefinitne surečenice (v. također Borucinsky, 2015). Računalnojezikoslovni alati, unatoč napretku tehnologije, ni dan-danas ne rade sa stopostotnom točnošću, a ona je moguća „samo ako nakon ili za vrijeme obrade slijedi ljudska intervencija u tekst ili ispravljanje pogrešaka“ (Bekavac 2002: 173).

„Strojno obilježavanje tekstova nekoga jezika iznimno je složen zadatak: bilo sa stajališta opsega posla koji treba obaviti, interdisciplinarnosti koje ono zahtijeva (lingvistika, informatika i statistika) ili količine znanja/ljudi koji se strojnom obradom bave“ (Bekavac 2002: 174). Za hrvatski postoji više od 900 morfosintaktičkih oznaka, a točnost označivača seže od 86,05% (*CroTag*, Agić, Tadić i Dovedan, 2009), preko 92,53% (*Reldi-tagger*, Ljubešić i Erjavec, 2016) i 93,87% (CLASSLA, Ljubešić i Dobrovoljc, 2019) do 95,81% (Ljubešić i Lauc, 2021).

Od važnih računalnojezikoslovnih alata za hrvatski (te srpski i slovenski) valja izdvojiti i lematizator, POS označivač i parser koji su dostupni i preko mrežnoga sučelja[51], a primjer jezične anotacije prikazan je u tablici 3. U tablici se navodi primjer rečenice i način na koji navedeni sustav vrši označavanje na razini lematizacije, označavanja vrsta riječi (POS) i sintaktičkoga parsera.

Tablica 3: *Primjer lematizacije i POS označavanja u* CLARIN-*u.*

	Surface	Tags	Lemma	Entity	Paragraph	Sentence	Token	Start char	End char
1.	Primjena	Ncfsn	primjena	O	1	1	1	1	8
2.	metoda	Ncfpg	metoda	O	1	1	2	10	15
3.	korpusne	Agpfsgy	korpusni	O	1	1	3	17	24
4.	lingvistike	Ncfsg	lingvistika	O	1	1	4	26	36
5.	u	Sl	u	O	1	1	5	38	38
6.	jezikoslovnim	Agpnply	jezikoslovni	O	1	1	6	40	52
7.	istraživanjima	Ncnpl	istraživanje	O	1	1	7	54	67

[51] *CLARIN*.

Tablica 4: *Primjer sintaktičkog parsanja u* CLARIN-*u.*

	Surface	Tags	Lemma	Dep parse - gov / func	Paragraph	Sentence	Token	Start char	End char
1.	Primjena	Ncfsn	primjena	0 / root	1	1	1	1	8
2.	metoda	Ncfpg	metoda	1 / nmod	1	1	2	10	15
3.	korpusne	Agpfsgy	korpusni	4 / amod	1	1	3	17	24
4.	lingvistike	Ncfsg	lingvistika	2 / nmod	1	1	4	26	36
5.	u	Sl	u	7 / case	1	1	5	38	38
6.	jezikoslovnim	Agpnply	jezikoslovni	7 / amod	1	1	6	40	52
7.	istraživanjima	Ncnpl	istraživanje	4 / nmod	1	1	7	54	67

Dubinsko parsanje (engl. *deep parsing*) odnosi se na anotaciju univerzalnih ovisnosti (engl. *universal dependencies*, UD). Morfološki dio sastoji se od lematizacije, oznaka za vrste riječi te niza obilježja kojima se kodiraju gramatički i leksički odnosi. Sintaktički dio opisa temelji se ovisnosti i riječi. Svaka riječ, osim one koja je glavna riječ odnosno korijen (engl. *root*), ovisi o nekoj drugoj riječi u rečenici. Kao što je prikazano u tablici 3, svaka oznaka sastoji se od broja (*gov*) i kratice (*funkcija*). Primjerice, pridjevom *korpusna* (pojavnica 3) upravlja imenica *lingvistika* (pojavnica 4), a ovaj pridjev je modifikator. Model morfosintaktičkoga označavanja standardnoga hrvatskog jezika ugrađen je u alat *CLASSLA-StanfordNLP tool*[52] uvježbavanjem na korpusu *hr500k* (v. fusnotu 40), te sadrži *CLARIN.SI-embed.hr word embeddings* (*Word Embeddings CLARIN.SI-Embed.Hr 1.0*). Model istovremeno koristi UPOS (*universal* POS), FEATS i *treebank-specific* POS (XPOS) (MULTEXT-East) oznake, a očekivana mjera preciznosti (F1) označavanja XPOS-a procjenjuje se na ≈94.1 (usp. Agić i Ljubešić, 2015; Samardžić i dr., 2017).

Višemilijunski korpusi ne mogu se ručno označavati, a čak se i označavanje manjih korpusa provodi (polu)automatski i ovisi o tzv. zlatnom standardu obilježavanja, iako i zlatni standardi označavanja vrsta riječi ili sintaktičkih obilježja koji su trenutno dostupni sadrže značajan broj pogrešaka (Dickinson i Meurers, 2003, 2005; Květoň i Oliva, 2002). Ne treba podcjenjivati učinak čak i nekoliko postotaka pogrešaka u označavanju korpusa, posebice kada se u obzir uzme Zipfov zakon (v. 3.). Dok se većina jezikoslovaca slaže oko načina na koji se vrše lematizacija i označavanje vrsta riječi, to nije slučaj kod označavanja sintaktičkih kategorija, tj. kod parsera. Naime, POS označavanje je u pravilu neproblematično, a vrsta riječi se u većini slučajeva može odrediti iz položaja riječi u rečenici, iako, naravno postoje i iznimke, kao što je kasnije

[52] GitHub - CLARINSI/CLASSLA.

pokazano na primjeru oznake *Xf*. Za razliku od toga parser mora krenuti iz neke sintaktičke teorije, pa se tako primjerice unutar teorije Univerzalne ovisnosti (engl. *Universal Dependency*) postavlja okvir za sustavnu gramatičku i sintaktičku anotaciju koja se može primijeniti na razne jezike. Osnovna reprezentacija ovisnosti prikazuje se u obliku stabla pri čemu je jedna riječ uvijek glava rečenice, a ovisi o korijenu, dok su ostale riječi ovisne o drugim riječima u rečenici (v. tablice 3 i 4). Sintaktičko je označavanje, dakle, rezultat pojedine gramatičke teorije ili teorijskoga pristupa prema kojemu se izvodilo, bez obzira na to što bi ono u biti trebalo biti neutralno ili što neutralnije. Budući da sintaktički pristupi u hrvatskome jezikoslovlju nisu u potpunosti razrađeni i da se mnogi miješaju, posljedica je nepravilno morfosintaktičko označavanje. To je vrlo ozbiljan problem koji u pitanje dovodi rezultate pretraživanja prema morfosintaktičkim kategorijama. Nadalje, promatrani sintaktički obrazac ili kategorija (v.7.2) može lako imati samo nekoliko pojavljivanja u korpusu, a ako je k tome i pogrešno označen, preciznost upita i dobiveni rezultati neće biti valjani. Nadalje, pogrešno označeni dijelovi korpusa nisu ravnomjerno raspoređeni po tekstovima, što također može utjecati na dobivene rezultate pa valja biti oprezan sa zaključcima i intepretacijom podataka dobivenih iz korpusa.

Označavanje nije ograničeno samo na vrste riječi ili sintaktička obilježja. Bilo koji jezični aspekt može se označiti u korpusu, jedino je pitanje može li se označavanje izvesti automatski ili ne. Formalna jezična obilježja moguće je automatski dodijeliti jer se na temelju gramatičkih pravila jezika pomoću probabilističkih modela određuje vjerojatnost da neka riječ pripada određenoj kategoriji. Za razliku od toga, funkcionalna obilježja (kao npr. upravni i neupravni govor) ne mogu se automatski dodijeliti. No, vrijednost obilježavanja leži u činjenici da omogućuje kvantifikaciju drugih jezičnih značajki kao što su prozodija i vrsta surečenice, koje nam mogu dati uvid u to kako jezik kao sustav funkcionira (Meuers i Müller, 2009).

Zaključno, obilježeni korpus omogućuje ponavljanje korpusno utemeljenih ili korpusom vođenih istraživanja, čime korpus postaje višefunkcionalnim jer se kao metodološki konstrukt može koristiti u leksikografiji, strojnome prevođenju, poučavanju jezika, analizi diskursa i mnogim drugim potpodručjima.

3. TERMINOLOGIJA

U ovome poglavlju prikazat ćemo terminologiju[53] koja će se rabiti u nastavku knjige te pojasniti pojmove koji su od važnosti da bi se mogao pratiti daljnji slijed knjige. Razlog zbog kojeg pojmovi nisu svrstani u glosar na kraju knjige jest taj što ih je potrebno pobliže opisati te ilustrirati primjerima, formulama i slikama. Pojmovi nisu posloženi abecednim redoslijedom već se polazi od najosnovnijih pojmova i nastavlja ka složenijima.

Jedan od temeljnih pojmova u korpusnolingvističkim istraživanjima jest **pojavnica** (engl. *token, running word*). Iako se gdjekad poistovjećuje s riječju, ona nije riječ, već najmanja jedinica od koje je korpus sastavljen, ili sve što se nalazi između dvije bjeline. Pojavnica tako može biti i riječ ali i neriječ, odnosno pojavnica koja ne započinje slovom i sadrži interpunkcijski znak, kao npr. *25-satni*.[54] Primjerice, *format A1* dvije su pojavnice, kao i riječ *nemo'š*. Potonja sadrži interpunkcijski znak, a opcija *glue* u *SkE-u* (v. 2.) takve riječi spaja pa ih prikazuje kao jednu cjelinu. Kao što je poznato iz raznih jezikoslovnih rasprava, riječ možemo definirati iz različitih perspektiva pa tako postoji fonološka riječ, ortografska riječ, morfosintaktička riječ, leksička riječ i dr.[55] Stoga ne čudi činjenica da ni pojavnica nema univerzalnu definiciju, a različiti alati programirani su da na određen način broje pojavnice. U *SkE*-u u pojavnice se ubrajaju i interpunkcijski znakovi, dok se u nekom drugom alatu kao što je *#LancsBox* ne ubrajaju, stoga je važno da korisnik točno zna kako računalnojezikoslovni alat kojim se služi funkcionira. Primjerice, *#LancsBox* u isječku pod (1) broji 9 pojavnica, dok *Word calculator* iz *Lancaster Stats Tool Online (LSTO)*[56] broji 10 pojavnica, a *SkE* 11 pojavnica.

(1) Ovo je samo test.
nemo'š
A1 format
računalno-jezikoslovni alat

Te razlike mogu se objasniti činjenicom da *#LancsBox* i *LSTO* ne broje interpunkcijske znakove kao pojavnice, dok ih *SkE* broji.

[53] Terminologija opisana u ovome poglavlju pokriva potrebe rada u računalnojezikoslovnom alatu *SkE*, ali obuhvaća i pojašnjenje procesa koji se događaju kada računalu postavimo upit. Detaljne upute dostupne su na *User Guide* | *Sketch Engine*.

[54] Regularni izraz (v. Tablicu 2) za pronalaženje neriječi u SKE je [^[:alpha:]].* .

[55] Riječ | *Hrvatska Enciklopedija*

[56] *Statistics in Corpus Linguistics: Lancaster Stats Tools Online*

Različnica je jedinstven oblik pojavnice (engl. *type, unique token*), odnosno pojedinačna riječ koja se razlikuje od druge ili drugih riječi.[57] U korpusu je to riječ koja se bilježi samo pri prvome pojavljivanju jer se sa svakim sljedećim pojavljivanjem smatra pojavnicom.

Osim pojavnica i različnica, pojam **leme** odnosno natuknice bitan je za postavljanje upita pomoću jezika za postavljanje upita u korpusu (*corpus query language*, CQL). Lema je pozicijski atribut, a definira se kao osnovni, natuknički ili rječnički oblik. Leme su specifične za svaki jezik ali i za pojedne računalnojezikoslovne alate, pa su tako *many, more, most* u *SKE*-u tri leme u engleskome jeziku, iako se gramatički gledano mogu svesti na jednu zajedničku lemu *many.* Pojedini autori uvode i pojam **leksema** kako bi razlikovali homonime, tj. leme koje imaju isti oblik a različito značenje, kao npr. vila:[58] 1. mlado žensko biće s magičnim moćima, 2. raskošni ljetnikovac obično okružen parkom.[59]

Broj i omjer pojavnica, različnica i lema bitan je, između ostaloga, za istraživanja vokabulara i poučavanja jezika, te u analizi diskursa. Ti omjeri pokazuju koliko je bogat ili siromašan tekst, te koliko ga se lako može čitati; ukratko, njima se mjere struktura i složenost teksta, što također može biti značajno za proučavanje prijevoda odnosno prevedenih tekstova u odnosu na izvorne tekstove.

Leksička gustoća (engl. *lexical density*) pokazuje omjer leksičkih riječi i/ili gramatičkih riječi u odnosu na ukupan broj riječi ili rečenica u korpusu. Leksička gustoća može se izraziti kao:

$$leksička\ gustoća = \frac{broj\ leksičkih\ ili\ gramatičkih\ riječi}{ukupan\ broj\ riječi}\ x\ 100\ (1.1)$$

Leksička raznolikost, za koju se u engleskome rabe nazivi *lexical diversity* i *lexical variation* također je jedan od pokazatelja leksičkog bogatstva ili siromaštva jezika. Jedna od mjera leksičke raznolikosti jest **omjer različnica i pojavnica** (engl. *type/token ratio*) u korpusu, što se može izraziti na sljedeći način:

---

[57] *Pojmovnik – Hrvatski mrežni rječnik*.

[58] *Hrvatski jezični portal.*

[59] Podatke o broju pojavnica, različnica, lema i dr. moguće je u *SkE*-u iščitati u sučelju *hrWaC Corpus Info*.

$$omjer\ različnica\ i\ pojavnica = \frac{broj\ različnica\ u\ tekstu\ ili\ korpusu}{broj\ pojavnica\ u\ tekstu\ ili\ korpusu} \quad (1.2)$$

Što je veći broj različnica u tekstu, tekst ima veću leksičku raznolikost (Březina 2018: 57). Usporedimo li tekstove A i B iz tablice 5, i unesemo li podatke u formulu (1.2), vidimo da je omjer pojavnica i različnica u tekstu A 0,53, a u tekstu B 0,87. Prema tome, a kao što možemo i pretpostaviti s obzirom na to da tekst A pripada razgovornome stilu, a tekst B znanstvenome, tekst B ima veću leksičku raznolikost, odnosno bogatiji je i složeniji.

Tablica 5: *Primjer teksta za ispitivanje strukture i složenosti*

Tekst A	**Tekst B**
BAU: ja jučer doša na sat počela kiša. *BAT: xxx kiša.* *BAT: je i o(v)de.* *BAT: je li ovde bila kiša?* *BAU: udria pljusak.* *BAT: ovde je bila neka mala kiša.* *BAU: e: ka(o).* *BAT: ka(ž)en ti asu ja pa bilo je vruće sidit.* *BAT: eto ja* *BAT: virujen ti ja.* *BAT: bia san u onoj tankoj majčici bilo mi je vruće sidit.*[60]	*Jezikoslovlje se sastoji od niza pojedinačnih disciplina, već prema tomu kojim se jezičnim razinama bave (fonetika i fonologija, uključujući i prozodiju, zatim grafemika, monematika /morfematika/ i morfologija, sintaksa, tvorba riječi, leksikologija, semantika itd.) ili s kojega gledišta jeziku pristupaju (psiholingvistika, sociolingvistika, pragmatika, tekstna lingvistika /lingvistika teksta/, kognitivna lingvistika i dr.).*[61]

Međutim, valja imati na umu da je leksička raznolikost osjetljiva na veličinu teksta i valjan rezultat dobit će se za tekstove koji sadrže sličan broj riječi, dok za tekstove koji se znatno razlikuju po broju riječi valja koristiti standardizirani omjer različnica i pojavnica (engl. *standardized type/token ratio*, Scott, 2008) ili prosječan omjer različnica i pojavnica (engl. *moving average type/token ratio* (Covington i McFall, 2010) koji tekst segmentiraju na manje isječke te potom računaju leksičku raznolikost.

[60] Tekst je preuzet iz korpusa govornoga jezika (*hrAL*), v. tablicu 1.

[61] Tekst je preuzet iz sljedećega izvora: *Jezikoslovlje* | *Hrvatska Enciklopedija*.

Osim omjera različnica i pojavnica, na isti način može se izračunati i omjer lema ili natuknica i pojavnica (Berman, 2008). Što je taj omjer veći, odnosno bliže vrijednosti 1, u tekstu je više različitih ili novih riječi: često sinonima, rjeđih i rijetkih riječi, stručnih riječi. Što je omjer manji odnosno bliže vrijednosti 0, u tekstovima se više riječi ponavlja. Kada se izražava postotcima, natuknice se množe sa 100 prije dijeljenja s brojem pojavnica (usp. također Jelaska i Baričević, 2012; Tomašić i Brkić, 2012).

Leksička raznolikost i leksička gustoća često se znaju poistovjećivati, ili se pak smatra da su vrijednosti proporcionalne, tj. ako je jedan omjer velik, bit će i drugi, što ne mora biti točno. Tekst, primjerice, može imati veliku leksičku raznolikost, odnosno velik broj različitih oblika riječi, ali malu leksičku gustoću ako su ti oblici uglavnom gramatičke riječi, ili pak obrnuto – tekst može imati malu leksičku raznolikost, a veliku gustoću, tj. velik broj leksičkih riječi kao što su imenice, pridjevi ili glagoli (Johansson, 2008).

Kada računamo frekvencije u korpusu, valja razlikovati **apsolutnu frekvenciju** (engl. *absolute frequency, raw frequency*) od **relativne ili normalizirane frekvencije** (engl. *relative, normalized frequency*). Apsolutna je frekvencija[62] zbroj svih pojavljivanja neke riječi u korpusu. Primjerice, riječ *oblak* pojavljuje se 3295 puta u *Riznici* i 36 463 puta u *hrWaC*-u, a riječ šerbet 6 puta u *Riznici* i 76 puta u hrWaC-u. Najčešće su riječi u nekom korpusu gramatičke riječi, a u hrvatskome je to veznik *i* s apsolutnom frekvencijom 44 826 297 u *hrWaC*-u te 2 989 033 u *Riznici*. Apsolutna frekvencija dobra je mjera ako promatramo samo jedan korpus. No, želimo li usporediti korpuse različitih veličina, vrijednosti je potrebno normalizirati na način da se ukupan broj pojavnica, tj. apsolutna frekvencija riječi, podijeli s ukupnim brojem riječi u korpusu te pomnoži s bazom za normalizaciju koja može biti 10 000 ili čak 1 000 za male korpuse, no najčešće je 1 000 000. Relativnu frekvenciju dobit ćemo na sljedeći način:

$$relativna\ frekvencija = \frac{apsolutna\ frekvencija}{broj\ pojavnica\ u\ korpusu}\ x\ norm.baza \quad (1.3)$$

Relativnu frekvenciju za riječ *oblak* u oba opća korpusa hrvatskoga jezika izračunat ćemo, dakle, na sljedeći način:

$$relativna\ frekvencija\ (oblak_{Riznica}) = \frac{3295}{101782863}\ x\ 1000000 = 32{,}4 \quad (1.4)$$

$$relativna\ frekvencija\ (oblak_{hrWaC}) = \frac{36463}{1405794913}\ x\ 1000000 = 25{,}9 \quad (1.5)$$

[62] Osim naziva frekvencija, moguće je rabiti i nazive *čestoća, čestotnost i frekventnost.*

Riječ *oblak* u korpusu *hrWaC* pojavljuje se 25,9 puta na milijun pojavnica i manje je česta u odnosu na istu riječ u korpusu *Riznica*. Unesemo li na isti način brojeve u formulu (1.3), utvrdit ćemo da je riječ šerbet neznatno češća u *Riznici* s relativnom frekvencijom od 0,06, dok je njezina relativna frekvencija u *hrWaC*-u 0,05.

Računalnojezikoslovni alati automatski će izračunati relativnu frekvenciju. Podaci koji se mogu dobiti iz *SkE*-a prikazani su na slici 4, a interpretirati ih možemo na sljedeći način:

Riječ *oblak* pojavljuje se 36 463 puta u korpusu *hrWac*, odnosno 25,9 puta na milijun pojavnica, te čini 0,002594 % cjelokupnoga korpusa.

Relativna frekvencija korisniku korpusa olakšava iščitavanje frekvencija, no valja imati na umu da odabrana baza za normalizaciju može utjecati na interpretaciju rezultata. Primjerice, ako uzmemo bazu koja je prevelika proporcije će biti matematički točne, no korisnik korpusa neće točno iščitati rezultate (Březina 2018: 43). Drugim riječima, za male korpuse ne smijemo uzeti bazu veću od broja riječi u korpusu. Osim apsolutne i relativne frekvencije postoji niz mjera koje će biti pogodne za tekst ili korpus, a ovdje ćemo ukratko pojasniti **srednju reduciranu frekvenciju** (engl. *average reduced frequency*, ARF, Hlaváčová, 2006; Savický i Hlaváčová, 2002) koja predstavlja važnu mjeru za sastavljanje rječnika i enciklopedija te za određivanje N-grama koji se mogu smatrati učestalim leksičkim spojevima (engl. *lexical bundles*, v. 7.1.4.). ARF je mjera kojom se kombiniraju frekvencija i disperzija. Može se, primjerice, dogoditi da se neka riječ često pojavljuje u jednome tekstu, ali da nije (jednako) zastupljena u svim tekstovima. Stoga je uvedena dodatna mjera, a to je položaj na kojem se riječ nalazi u korpusu. Uzmimo da se riječ r_1 pojavljuje 5 puta u jednomilijunskome korpusu, a riječ r_2 također pet puta u jednomilijunskome korpusu. Obje dakle imaju istu apsolutnu i relativnu frekvenciju, no prva riječ r_1 pojavljuje se samo u jednome tekstu, dok je druga r_2 ravnomjerno raspoređena u svim tekstovima. Druga riječ prema tome ime veću srednju reduciranu frekvenciju i bit će značajnija prilikom donošenja odluke o njezinu uvrštavanju u rječnik, enciklopediju i sl.[63] Ovom se mjerom ukida neravnomjernost riječi u korpusu koja može negativno utjecati na rezultat, a može biti i vrlo korisna u izradi frekvencijskoga ili čestotnoga rječnika.

Disperzija (engl. *dispersion*) pokazuje u kojoj su mjeri riječi rasprostranjene u korpusu. Primjerice, veznik *i* koji je najčešća riječ u korpusu hrvatskoga jezika

[63] ARF se može izračunati automatski u *SkE*-u ili u *LSTO-u.*

ravnomjerno je raspoređen u svim vrstama tekstova tj. potkorpusima *Riznice* (slika 8), dok se riječ *prijestolonasljednik* pojavljuje u samo dva od četiri potkorpusa *Riznice* (slika 9).

	Subcorpus	Frequency	Relative density
1	Vjesnik	2,156,691	97.46
2	Knjige	665,557	115.68
3	Sportske	96,003	73.47
4	Zakoni	71,008	101.11

Slika 8: *Disperzija veznika* i *u* Riznici

	Subcorpus	Frequency	Relative density
1	Vjesnik	81	109.42
2	Knjige	19	98.72

Slika 9: *Disperzija riječi* prijestolonasljednik *u* Riznici

Veznik *i* pojavljuje se u svim potkorpusima jer se radi o gramatičkoj riječi bez koje bismo teško mogli slagati rečenice ili sintagme, dok su riječi kao što je *prijestolonasljednik* specifične za određeni kontekst pa je i manja vjerojatnost da će se pojaviti u tekstovima iz tematskoga područja sporta ili u zakonodavnim tekstovima. Postoje različite mjere disperzije riječi u korpusu, kao što su raspon (engl. *range*), standardna devijacija (engl. *standard deviation*), standardna devijacija uzorka (engl. *sample standard variation*), koeficijent varijacije (engl. *coefficient of variation*), Juilliandov D, devijacija proporcija (engl. *Deviation of Proportions,* DP) (usp. Březina 2018).

Kod primjene korpusa u jezikoslovnim istraživanjima važno je imati na umu da je korpus (pomno) sastavljen jezični uzorak i da nam pokazuje samo ono što se u njemu nalazi, što predstavlja svojevrsno ograničenje. Stoga valja imati na umu da količina dokaza koje možemo dobiti iz korpusa naglo pada. Ta se pojava naziva **Zipfovim zakonom**, a opisuje princip naglog opadanja broja različnica u korpusu, odnosno pokazuje da će druga najfrekventnija riječ imati dvostruko manju frekvenciju od prve, treća samo trećinu itd. Stoga se može dogoditi da u korpusu ne možemo pronaći potvrdu ili dokaz za jezičnu pojavu koju istražujemo, što ne znači da je bez kritičkog promišljanja možemo odmah odbaciti kao nepostojeću.

Kolokacija u korpusnoj lingvistici označava sustavno supojavljivanje riječi u uporabi, odnosno riječi koje se češće pojavljuju jedna uz drugu, i kao posljedica toga međusobno utječu na značenje. Postoji nekoliko statističkih mjera kojima se određuje jačina (su)pojavljivanja, a dobiveni rezultati često su

specifični za korpus i ne mogu se uvijek rabiti za usporedbu kolokacijske sveze u različitim korpusima (Rychlý, 2008).

Mjera međusobne informacije (engl. *mutual information*, MI) jedna je od statističkih mjera za izračun povezanosti dviju riječi, a računa se tako da se podijeli opažena frekvencija (engl. *observed frequency*) kolokata u definiranom rasponu za osnovu (kolokator[64]) s očekivanom frekvencijom (engl. *expected frequency*) kolokata u istom rasponu, te izvede logaritam rezultata (Xiao, 2015: 109), kao što je prikazano u formuli (1.6)[65].

$$MI = log_2 \frac{f_{AB}N}{f_A f_B} \quad (1.6)$$

gdje N označava ukupan broj riječi u korpusu (npr. 85 273 724 riječi u korpusu *Riznica*), je broj supojavljivanja riječi A i B u korpusu u zadanome rasponu (npr. 3 riječi lijevo i desno od kolokatora, gdje je raspon (R = 6)), je broj pojavljivanja riječi A (npr. *siv(i)* ima frekvenciju 3151), a broj pojavljivanja riječi B (npr. *oblak* ima frekvenciju 3295). Log2 je konstanta i iznosi približno 0,301. Prema tome mjera međusobne informacije kolokacije *sivi oblak* u *Riznici* iznosi 10,93. Valja imati na umu da veći raspon daje veći MI. Ovo je mjera asocijativne snage koja pokazuje jesu li dvije leksičke sastavnice povezane. Što je taj broj veći, veća je povezanost među riječima, a što je bliži 0 (ili ako je negativan) veća je vjerojatnost da se riječi slučajno supojavljuju. U pravilu se postavlja prag od 3,0 kao statistički značajna mjera međusobne informacije za određivanje da dvije riječi koje se supojavljuju predstavljaju značajnu kolokaciju i da njihovo supojavljivanje nije nasumično (Hunston 2002: 71). Mjera MI nije pouzdana za kolokacije čije sastavnice imaju velika odstupanja u frekvenciji, stoga, kako navodi Hunston (2002: 72), nije uvijek pouzdana u određivanju značajnosti, pa valja uzeti u obzir i veličinu korpusa, za što je korisna tzv. **T-mjera** (engl. *T-score*), koja se temelji na aritmetičkoj sredini i varijanci uzorka koji se uspoređuje unutar očekivane srednje vrijednosti kada je potvrđena nulta hipoteza[66]. Mjera se dobiva na temelju razlike između aritmetičke sredine opažene frekvencije i aritmetičke sredine očekivane frekvencije, kako

[64] Osim naziva baza koji je u leksikografiju uveo Hausmann (2007), u hrvatskoj se literaturi rabe i nazivi osnova i ključna riječ (Pritchard, 1998), dok se u korpusnoj lingvistici rabi naziv *node* (dosl. prijevod *čvor*).

[65] Kilgarriff i dr. (2015a).

[66] Nulta hipoteza pretpostavlja da ne postoji statistički bitna razlika između ispitivanih struktura (Izvor: *Struna*).

bi se utvrdila vjerojatnost uzorka sredine i varijance s pretpostavkom da skupina podataka ima normalnu distribuciju, a može se izraziti na sljedeći način:

$$T - mjera = \frac{f_{AB} - \frac{f_y f_B}{N}}{\sqrt{f_{AB}}} \quad (1.7)$$

Uvrstimo li podatke o kolokaciji *sivi oblak* u formulu (1.7), dobit ćemo 4,12, a ova mjera pokazuje da je nasumičnost supojavljivanja dviju riječi potkrijepljena značajnom količinom dokaza u korpusu (Evert, 2005, usp. također Posavec, 2017). T-mjera od 2,576 ili iznad smatra se statistički značajnom, što znači da je pridjev *sivi* značajan kolokat osnove *oblak*. Dok mjera MI mjeri jačinu kolokacije, T-mjera mjeri pouzdanost s kojom možemo tvrditi da postoji povezanost među riječima (Church i Hanks, 1989). Osim ovih dviju mjera, pokazatelj asocijativne snage među riječima je i **Diceov koeficijent** (Dice, 1945) koji dijeli presjek skupova i sa zbrojem skupova i , a rezultat normalizira faktorom 2[67].

$$D = \frac{2f_{AB}}{f_A + f_B} \quad (1.8)$$

Ova mjera smatra se prikladnom za pokazivanje povezanosti među riječima, no kako je rezultat obično izražen vrlo malim brojevima, definiran je **logDice** (Rychlý, 2008) da se doskoči problemu, jer je takve vrijednosti teško iščitavati i uspoređivati.

$$logDice = 14 + \log_2 D = 14 + \log_2 D = \frac{2f_{AB}}{f_A + f_B} \quad (1.9)$$

U primjeru *sivi oblak logDice* iznosi 7,287 što znači da postoji statistički značajna povezanost među riječima *sivi* i *oblak* te da ih se može smatrati kolokacijom. Kada se sve pojavnice riječi A supojavljuju s pojavnicama riječi B i obrnuto, tada je rezultat 14, što je teoretski maksimalan iznos. Rezultat 0 pokazuje da postoji manje od jednog supojavljivanja riječi A i B na 16 000 pojavnica, dok negativna vrijednost znači da ne postoji statistički značajna povezanost među riječima, pa ih se ne smatra kolokacijama.

[67] Usp. također Posavec (2017).

Log izglednost (engl. *log likelihood*) jedna je od mjera povezanosti, a temelji se na funkciji izglednosti kojom se određuje statistička značajnost. Može se izračunati na način da se postavi sljedeća tablica kontingencije:[68]

Tablica 6: *Tablica kontingencije*

	Korpus 1	**Korpus 2**	**Ukupno**
Frekvencija riječi	a	b	a+b
Frekvencija ostalih riječi	c-a	d-b	c+d-a-b
Ukupno	c	d	c+d

Vrijednost *c* predstavlja broj riječi u korpusu 1, a *d* broj riječi u korpusu 2, što su dvije N vrijednosti. Vrijednosti *a* i *b* su opažene vrijednosti, a očekivane vrijednosti dobit će se pomoću sljedeće formule:

$$E_i = \frac{N_i \sum_i \quad O_i}{\sum_i \quad Ni} \quad (1.10)$$

što znači da je N1 = c, a N2= d, iz čega slijedi:

$$E = c*(a+b)/(c+d);\ E2= d*(a+b)/(c+d) \quad (1.11)$$

pri čemu se u izračunu uzima u obzir veličina korpusa pa stoga nije potrebno normalizirati vrijednosti prije uvrštavanja u formulu. Na temelju dobivenih vrijednosti za opaženu i očekivanu frekvenciju možemo izračunati log izglednost (Rayson i Garside, 2000): [69]

$$-2ln\lambda = 2\sum_i \quad O_i \; ln \; ln \left(\frac{O_i}{E_i}\right) \quad (1.12)$$

Log izglednost omogućuje da se statistički usporede dva korpusa te odredi postoji li među njima statistički značajna razlika. Kada je broj koji se dobije kao rezultat veći od 3,84, vrijednost *p* je manja od 0,05 (značajnost), pa se može utvrditi da je razlika statistički značajna.

Konkordancija[70] (engl. *concordance*) jest popis riječi sa svim oblicima koji se nalaze u korpusu zajedno s njihovim ko(n)tekstom (engl. *keyword in context*

[68] Tablica kontingencije je tablica koja u recima i stupcima sadrži frekvencije atributivnih obilježja. Opažanja su klasificirana po više atributa ili opažanja iz više uzoraka mogu biti klasificirana po kategorijama jednoga atributa. – nejasna mi je ova rečenica, provjeriti je li sve u redu?

[69] Više o log izglednosti v. Cumming (2014); Johnston i dr., (2006); Kilgarriff (2005); Kühberger i dr. (2014); Leek i Peng (2015); Lijffijt i dr. (2016); Purnelle i dr. (2004); Rayson (2008).

[70] Osnovne funkcionalnosti i način pretrage hrvatskih korpusa opisala je Posavec (2017).

(KWIC)).[71] Program koji omogućuje prikazivanje velikog broja riječi abecednim redoslijedom s kontekstom pojavljivanja u korpusu naziva se konkordancer.

Funkcionalnost **slučajni uzorak** (engl. *Get a random sample*) korisna je leksikografima jer omogućuje uočavanje tipične uporabe riječi te kod velikog broja rezultata smanjuje broj konkordancija koje je potrebno pregledati, a još uvijek čuva reprezentativnost konkordancija. Pri tome korisnik sam može definirati koliko nasumičnih primjera želi dobiti, a konkordancije će biti prikazane onim redoslijedom kojim se pojavljuju u korpusu. I najvažnije, slučajni će uzorak uvijek pokazivati iste konkordancije, čime se omogućuje ponovljivost danoga istraživanja koje se provodi ovom metodom (primjerice dva različita leksikografa vidjet će iste konkordancije ili studenti mogu pratiti iste korake na nastavi kako bi dobili isti rezultat ako se korpus koristi u nastavi[72]).

Skica riječi (engl. *word sketch*) prikazuje kolokacije kategorizirane prema gramatičkim odnosima, a da bi bila funkcionalna korpus mora biti označen prema vrstama riječi i za njega mora biti definirana gramatika za izradu skica riječi. Ova funkcionalnost dostupna je u *SkE*-u, no nije dostupna u besplatnoj inačici *NoSkE.*

Gramatika za izradu skica riječi (engl. *word sketch grammar*, *WSG*) niz je pravila koja definiraju gramatičke odnose, a specifična su za svaki jezik. Korisnici za potrebe istraživanja mogu definirati gramatiku.

Gramatika za crpljenje naziva (engl. *term grammar*) odnosi se na definiranu leksičku strukturu za svaki jezik na temelju koje se crpe nazivi. U pravilu obuhvaća imenske skupine[73], no može se proširiti i na glagolske skupine i dr.

Razlika u skicama riječi (engl. *word sketch difference*) prikazuje razlike među zadanim kolokacijama kategoriziranim prema gramatičkim odnosima, a da bi bila funkcionalna korpus mora biti označen prema vrstama riječi. Ova funkcionalnost dostupna je u *SkE*-u, no nije dostupna u besplatnoj inačici *NoSkE***.**

Tezaurus (engl. *thesaurus*) je funkcionalnost koja prikazuje sinonime i riječi koje imaju slično značenje, a temelji se na distribucijskoj semantici (v. 7.1.3.).

---

[71] *Pojmovnik - Hrvatski mrežni rječnik*.

[72] U ovoj knjizi nije posvećeno puno mjesta uporabi korpusa u nastavi. Pregledavanje konkordancijskih nizova težak je i nemotivirajući zadatak, posebice za studente. No, usprkos tome, korpusi se mogu rabiti u nastavi na način da se konkordancije prikaže kao rječnik (v. Borucinsky i Tominac Coslovich, 2021; Kilgarriff, 2009).

[73] U hrvatskim gramatikama prednost se daje nazivu imenska sintagma. Borucinsky (2015) po uzoru na model sistemske funckionalne gramatike predložila je naziv imenska skupina. Sintaktički opis imenske skupine u hrvatskome nije detaljno razrađen, niti temeljen na korpusnim istraživanjima (v. 7.2.).

Lista riječi (engl. *word list*) ili **frekvencijska lista, čestotna lista**, funkcionalnost je koja ispisuje popise imenica, glagola, pridjeva i dr. prema učestalosti pojavljivanja u korpusu.

N-grami predstavljaju znakove u nizu. U *SkE*-u to su riječi, a osim riječi u nizu, pod N-gramima se podrazumijevaju i slova u nizu, koja se obično nazivaju *character grams, char-grams* (Fletcher, 2012, v. 6.2.). U *SkE*-u opcija N-gram izlistava popis višerječnih naziva prema učestalosti pojavljivanja u korpusu. N-grami su sastavljeni od pojavnica, pa je tako izraz 'don't like' (v. također *glue*) trigram. N-grami su vrsta višerječnih izraza (engl. *multi-word expressions*), koji su u anglističkoj tradiciji poznati kao *lexical bundles* (Biber i dr., 1999), *clusters* (Scott, 2008), *chains* (Stubbs, 2001; Stubbs i Barth, 2003), *recurrent sequences* (De Cock, 2004), *recurrent word combinations* (Altenberg, 1998), i koji imaju važnu ulogu u oblikovanju diskursa i razlikovanja funkcionalnih stilova. Svaka kolokacija je N-gram, ali nije svaki N-gram kolokacija. *SkE* ima tehničko ograničenje te u iznimno velikim korpusima generira N-grame samo iz prvih milijardu pojavnica. No, moguće je generirati N-grame iz cijelog korpusa uz naknadu.

Ključna riječ (engl. *keyword*) je riječ koja ima veću frekvenciju u 'fokusnom' ili zadanom korpusu (engl. *focus corpus*) u odnosu na referentni korpus (engl. *reference corpus*), a može biti jednorječna (engl. *single word lexical unit*) ili višerječna (engl. *multi-word lexical unit,* MWLU). Potonja, osim uvjeta da je češća u fokusnom korpusu u odnosu na referentni korpus, mora zadovoljiti i uvjet da njezina struktura odgovara strukturi naziva koji su postavljeni za neki jezik (v. također *gramatika za crpljenje naziva*). Traženje ključnih riječi u korpusu od posebne je važnosti za specijalizirane korpuse koje istraživači u pravilu sami sastavljaju radi specifičnih potreba istraživanja. Riječi koje su češće u fokusnom korpusu u odnosu na referentni smatraju se pozitivnim ključnim riječima, one koje su češće u referentnome u odnosu na fokusni korpus negativne su ključne riječi, dok su tzv. *lockwords* riječi koje imaju sličnu frekvenciju u oba korpusa (Březina 2018: 80). **Ključnost** se riječi računa prema sljedećoj formuli (Kilgarriff, 2009):

$$ključnost = \frac{f\ pm_{rmfokus} + N}{f\ pm_{rmref} + N} \quad (1.13)$$

gdje predstavlja normaliziranu (na milijun, engl. *per million*) frekvenciju riječi u fokusnom korpusu, dok predstavlja normaliziranu (na milijun) frekvenciju riječi

u referentnom korpusu, a N je parametar kojim se izjednačavaju vrijednosti (obično se uzima broj 1).

Primjerice, riječ *cilindar* ima relativnu frekvenciju 2023,57 u korpusu *Brodostrojarstva* (veličina korpusa 1,2 milijuna pojavnica), u *hrWaC*-u 7,692 (veličina korpusa 1,9 milijardi pojavnica), što znači da je ključnost riječi *cilindar* 232,93, odnosno da se češće pojavljuje u jeziku brodostrojarske struke u odnosu na opći jezik. Na isti način možemo izračunati ključnost za višerječne nazive. Primjerice višerječni naziv *glavni motor* ima relativnu frekvenciju od 212,376 u specijaliziranom korpusu, a 0,290 u referentnom te je stoga njegova ključnost 165,38, dok *regulacija tlaka* ima frekvenciju od 14,558 u fokusnom korpusu i 0,1067 u referentnom korpusu s ključnošću od 14,058, što znači da se pojavljuje i u drugim specijaliziranim kontekstima (strojarstvu, medicini itd.).

Jezik za postavljanje upita korpusu (engl. *Corpus Query Language***)** „upit (je) koji omogućava pretragu korpusa po riječima, morfosintaktičkim oznakama ili pojedinim dijelovima riječi, oznaka ili izraza“ (Posavec 2017: 49). Pojam koji se pretražuje mora biti u navodnicima (" "), i mora mu biti dodijeljena vrijednost odnosno atribut (npr. riječ (*word*), lema (*lemma*), oznaka (*tag*), npr. [tag="Nc.*"] itd.).

Regularni izraz (engl. *regular expression*) niz je znakova s posebnom sintaksom koji se koriste za pronalaženje i manipuliranje tekstom (Forta, 2018). Oni čine zadani niz znakova koji se uspostavlja za pretraživanje korpusa pomoću alata *SkE* i *NoSkE* pri traženju ciljanih gramatičkih ili leksičkih uzoraka.[74] Regularni se izrazi koriste u *CQL*-u za definiranje uzoraka određenih vrijednosti (v. Tablicu 2).

[74] *Pojmovnik – Hrvatski mrežni rječnik*.

4. IZRADA KORPUSA

U poglavlju 1.7 utvrđeno je da za hrvatski jezik postoji relativno mali broj računalnojezikoslovnih resursa i alata. Proučavamo li opći (i standardni) jezik, na raspolaganju su nam *HNK*, *hrWaC* i *Hrvatska jezična riznica* (*Riznica*), te najnoviji hrvatsko-engleski usporedni korpus *MaCoCu-hr*, iako nijedan od navedenih korpusa nije reprezentativan u pravome smislu riječi (v. 1.4). No, za potrebe pojedinih istraživanja često je potrebno sastaviti vlastiti korpus u kojem će se, osim o uravnoteženosti i veličini, voditi računa i o tome da korpus sadrži i neku vrstu anotacije (v. 2). U ovome dijelu opisani su načini sastavljanja korpusa (ručno i automatsko sastavljanje), kriteriji za sastavljanje specijaliziranoga korpusa iz domene pomorskoga prava, te problemi s kojima se istraživač može susresti prilikom izrade korpusa.

4.1. RUČNO SASTAVLJANJE KORPUSA

Ručno je prikupljanje i sastavljanje korpusa jeftin, ali dugotrajan način sastavljanja korpusa, koji je prihvatljiv za manje korpuse, a podrazumijeva pronalaženje tekstova (iz vlastite arhive ili s mreže), te kopiranje i lijepljenje istih u *.txt*, *.doc* ili neki drugi dokument.[75] Najveća je prednost ovog načina sastavljanja korpusa mogućnost praćenja kvalitete tekstova budući da ih sami prikupljamo, za razliku od automatskoga načina u kojemu se tekstovi povlače iz zadanih mrežnih stranica. Takav dokument se može postaviti u većinu računalnojezikoslovnih alata kao što su *#LancsBox* (Březina i dr., 2020)*, AntConc* (Anthony, 2019a)*, WordSmith Tools* (Scott, 2008), *Sketch Engine* (Kilgarriff i dr., 2004), *MonoConc* (Barlow, 1999), *ParaConc* (Athelstan, 2010) i dr. na način da se tekstovi učitaju i alat ih automatski obilježi ili anotira. Od navedenih alata jedino *SkE* podržava i hrvatski jezik, odnosno tekstu dodjeljuje morfosintaktičke oznake, oznake vrsta riječi itd. (v. 2) specifične za hrvatski jezik. Iz tog razloga se u nastavku ove knjige, u kojem su navedena razna korpusna istraživanja, primjeri i studije, uglavnom koristi *SkE.*

[75] Uz to je potrebno u zaseban *.xls* dokument spremiti metapodatke o tekstovima (v. 2.). Alat *SkE* podupire sljedeće formate teksta: *.csv*, *.doc*, *.docx*, *.htm*, *.html*, *.ods*, *.pdf*, *.tar.bz2*, *.tar.gz*, *.tei*, *.tgz*, *.tmx*, *.txt*, *.vert*, *.xlf*, *.xliff*, *.xml*, *.zip.*

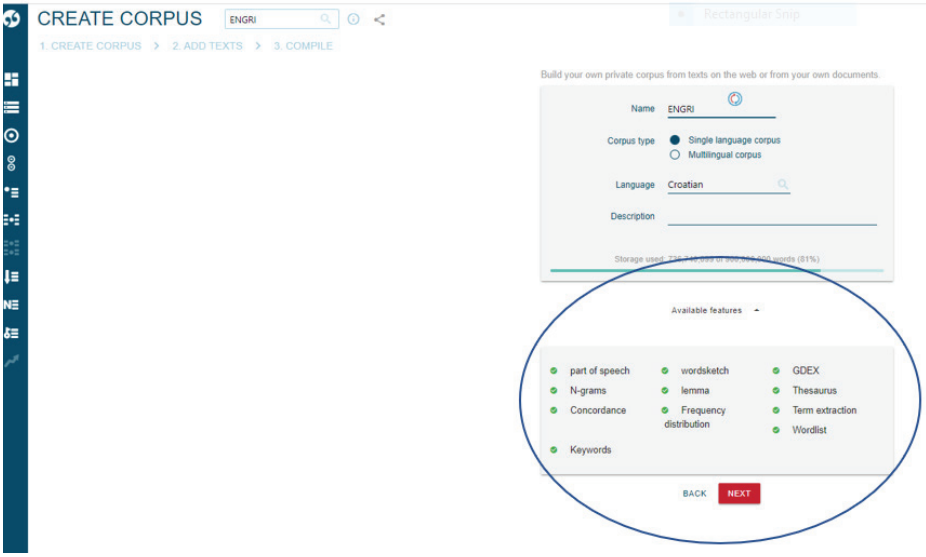
Prilikom učitavanja vlastitog teksta u odabrani alat ili sustav valja voditi računa o tome da tekst sadrži jedno pismo, da je pravilno kodiran, a formatiranje bi trebalo biti što jednostavnije pa se tako skenirane slike, primjerice, ne mogu učitavati. Nadalje, valja imati na umu da će alat učitati i tekst koji sadrži pravopisne pogreške, što može utjecati na rezultate dobivene korpusnom pretragom, posebice ako proučavamo pravopisne promjene u dijakronijskome korpusu.

4.2. AUTOMATSKO SASTAVLJANJE KORPUSA

Automatsko sastavljanje korpusa u *SkE*-u može se izvesti na dva načina: (1) povlačenjem URL-a ili mrežnih mjesta; (2) definiranjem riječi za pretragu (engl. *seed words*). Prvi način podrazumijeva da istraživač odabere i potom upiše točan URL ili mrežnu stranicu s koje će se crpiti jezični podaci. Primjerice, za korpus *ENGRI* prikupljeni su tekstovi s hrvatskih internetskih portala. Pri odabiru ove metode valja imati na umu da sve odabrane URL-ove ili mrežne stranice treba upisati ili kopirati jedne ispod drugih. Alat će potom pronaći navedene stranice te ukloniti nejezični sadržaj (npr. *HTML, boilerplate*). Ovaj način znatno je brži od ručnog prikupljanja korpusa, no sadrži određena ograničenja kao što su maksimalan broj stranica koje se mogu preuzeti (u *SkE*-u je to 2000), vrijeme preuzimanja od 6 stranica/min što je za velik korpus kao što je *ENGRI* vremenski vrlo neisplativo. Iz tog je razloga suradnik na projektu Hrvatske zaklade za znanost (2020.-2025.) *Engleske riječi u hrvatskome jeziku: identifikacija, afektivno-semantičko normiranje i ispitivanje kognitivne obrade bihevioralnim i neuroznanstvenim metodama* prikupio i pripremio tekstove s novinskih portala (usp. Kučić, 2021) te ih pohranio u relacijsku bazu *My SQL*. Iz baze je autorica preuzela korpus u *.xml* dokumentu koji je potom pretvorila u nekoliko manjih *.csv* dokumenata radi lakšeg i bržeg učitavanja u *SkE*. S obzirom na to da je korpus *ENGRI* prilično velik (v. tablicu 1), učitavanje je trajalo više sati uz povremene prekide i poteškoće kao što su pucanja veze, nepotpuno učitavanje, potrebe za povećanjem količine podataka dodijeljenih jednom korisniku itd. Nakon što je korpus učitan i nakon što su mu dodijeljena obilježja specifična za hrvatski jezik, podijeljen je sa svim zainteresiranim istraživačima. Moguće je takav korpus, uz dodatnu nadoplatu, učiniti dostupnim svim korisnicima koji posjeduju licencu za *SkE*.

Kao kod ručno prikupljenoga korpusa, i automatski prikupljeni tekstovi se tokeniziraju, lematiziraju, riječima se dodjeljuju oznake vrste riječi, morfosintaktička i druga obilježja (v. 2.). Neće svi jezici imati potpunu funkcionalnost u

SkE-u. Na slici 10 prikazane su funkcionalnosti koje su omogućene za hrvatski jezik (npr. *skica riječi*, *N-gram*, *thesuarus*, *GDEX* itd), a o kojima će više riječi biti u poglavlju 7.



Slika 10: *Prikaz dostupnih funkcionalnosti za hrvatski jezik u alatu* SkE

Automatsko sastavljanje korpusa vrši se pomoću tehnologije *WebBootCaT* (Baroni i dr., 2006), što rezultira tzv. *instantnim korpusom*. Budući da jezikoslovci u pravilu nisu informatičari i programeri, ova tehnologija osmišljena je upravo za takve korisnike, a temelji se na nekoliko jednostavnih principa. Prvi princip tiče se odabira riječi za unos ili tzv. *seed words,* drugi se odnosi na slanje upita i prikupljanje dobivenih podataka, što je moguće preko sučelja *WebBootCaT* ili *SkE*-a. S obzirom na to da sučelje *WebBootCaT* ne podržava hrvatski jezik, u nastavku će se prikazati kako je ista tehnologija implementirana u *SkE*-u.

Na temelju definiranih riječi *SkE* može sastaviti milijunski korpus u roku od 10 minuta. Kako bi sastavljeni korpus bio što bolji, potrebno je istražiti opcije kao što su *allowlist*, *denylist*, što je prikazano na slici 11, a detaljnije opisano u nastavku.

Slika 11: *Prikaz postavki automatskoga sastavljanja mrežnoga korpusa u alatu* SkE

Postavlja se, naravno, pitanje koje riječi ili fraze upisati u tražilicu. Primjerice, problematičan će biti korpus o gradu Rijeci u koji želimo uključiti povijesne činjenice i kulturne znamenitosti, ali ne želimo uključiti geografski pojam rijeke kao „većeg toka slatke vode koji teče koritom na površini Zemlje i ulijeva se u drugu rijeku, more ili jezero“ (HJP, 2021). Nadalje, pitanje je koliko riječi koristiti. Pravilo je da je 20 – 60 riječi dovoljno za određenu domenu. Mogu se upisivati i fraze i strane riječi, npr. *baby boom*, *all-inclusive ponuda*. Manje riječi rezultirat će manjim i fokusiranijim korpusom, no svakako je pri tome važno izbjegavati riječi koje možemo naći u drugim funkcionalnim stilovima ili područjima (npr. riječ *jezik* koja pripada domenama anatomije i lingvistike). Tehnologija potom kombinira riječi koje smo zadali u nasumične skupine od tri riječi i šalje ih tražilici *Bing* (ili *Google*), koja potom traži te riječi i vraća internetske stranice, ujedno uklanjajući reklame i navigacijske trake, odnosno neželjeni nejezični sadržaj. Riječi se mogu grupirati i u skupine od četiri riječi, čime se fokusira korpus, ali se dobije manje rezultata. *Bing* u pravilu pretražuje cijeli internet, no moguće je ograničiti pretragu na određene mrežne stranice ili domene. Određene mrežne stranice dopuštaju pristup samo nekim tražilicama, što znači da neće svi sadržaji biti dostupni *Bing*-u, a također se neće povući sadržaji koje tehnologija smatra jezično neprihvatljivima, kao što su vrlo kratki ili pak dugački tekstovi koji su podijeljeni u više manjih nepovezanih cjelina. To može biti problematično ako želimo prikupiti tekstove s foruma, u kojem bi slučaju metoda ručnoga sastavljanja korpusa bila bolja, ili je pak potrebno razviti algoritam koji će prikupiti takve tekstove (npr. *TweetCat* koji su razvili Ljubešić i dr. (2014)). I naposljetku, ostaje problem autorskih prava tekstova koji su prikupljeni s interneta, što je problem koji još valja riješiti.

Nakon što korisnik odobri pronađene internetske stranice, korpus se automatski kompilira. Potom je poželjno preko opcije crpljenja terminologije (engl. *extract terminology*) provjeriti sadržaj i prihvatljivost izrađenoga korpusa. Ovaj postupak sastavljanja korpusa može se ponavljati nekoliko puta kako bi se povećao i poboljšao korpus, a *SkE* će u pravilu sam ukloniti duplikate. Dok tehnologija *WebBootCaT* može ukloniti neželjeni nejezični sadržaj, razlikovanje kvalitete teksta programerski je vrlo zahtjevno, tako da je moguće da program ukloni i visokokvalitetan sadržaj. Nadalje, poželjno je postaviti minimalnu i maksimalnu veličinu datoteke kako se ne bi povlačili tekstovi koji se sastoje od jednoga odlomka, osim, naravno u slučaju da istražujemo isključivo kraće tekstove, odnosno tekstove foruma, blogova, objave s Twittera i sl.

Uz navedeno valja provjeriti valjanost podataka, što se posebno odnosi na transkripte i tekstove koji su preuzeti s interneta na način da se uklone oznake *html*, standardni kod (engl. *boilerplate*) i sl. Naposljetku, važno je odabrati odgovarajući alat za sastavljanje i, u konačnici, za analizu korpusa. Postojeći alati za sastavljanje i analizu korpusa opisani su u nastavku.

4.3. KRITERIJI ZA SASTAVLJANJE KORPUSA

U prvome poglavlju naveli smo da nije moguće istražiti jezik u cijelosti pa je za sastavljanje korpusa bitno da prikupimo uzorak odnosno dovoljnu količinu tekstova koji će adekvatno predstavljati jezik koji proučavamo. To prije svega znači da moramo odrediti koje tekstove uključiti u korpus. Svaki korpus, bez obzira na njegovu veličinu, sastoji se od pojedinačnih tekstova, pri čemu neki mogu biti duži (knjige, disertacije, zakoni itd.), dok se drugi mogu sastojati od svega nekoliko riječi (e-poruke, obavijesti i sl.). Sastavljanje korpusa moguće je provesti ručno ili automatski, no prije toga istraživači trebaju promisliti o nekoliko bitnih pitanja. Prvo se pitanje ili odluka koju valja donijeti odnosi na način izrade korpusa, te jezik ili jezični varijetet koji korpus kao uzorak treba predstavljati. U procesu izrade korpusa bitno je da istraživači vode bilješke o tome koje su tekstove uvrstili u korpus, a koje nisu, te razloge za potonje. Ako se korpus sastavlja ručno (v. 4.1.), poželjno je tekstove spremati kao zasebne dokumente tako da je kasnije omogućena jezična pretraga i prema vrsti tekstova u raznim dijelovima korpusa.

Sukladno preporukama projekta Europske unije EAGLES[76], koji obuhvaća i obrađuje problematiku sastavljanja i računalne podrške korpusima te predlaže standarde za njihovo kodiranje i obradbu (usp. Tadić, 1997), kriteriji za sastavljanje korpusa mogu biti vanjski i unutarnji, odnosno nejezični i jezični, a potonji se odnose na razlikovna svojstva tekstova. Kriteriji pritom nisu isključivi, već utječu jedan na drugi, pa je stoga važno uskladiti oba kriterija. Jezične kriterije valja provjeriti kroz prizmu vanjskih kriterija te obje kategorije prilagoditi, a postupak je potrebno ponavljati dok se ne postigne stabilnost uzorka (usp. Biber, 1993). Postupak uzorkovanja u pravilu započinje odabirom vanjskih kriterija, a potom slijedi fina raščlamba prema unutarnjim kriterijima. Oslanjanje isključivo na vanjske kriterije pri odabiru tekstova moglo bi dovesti do zanemarivanja značajnih varijacija među tekstovima, dok odabir vođen isključivo unutarnjim

[76] *MULTEXT-East Croatian part of speech tagset*.

kriterijima ne bi pružio podatke o odnosu između teksta i konteksta. Vanjski kriteriji definirani su situacijski, društveno, izvanjezično, bez obzira na distribuciju jezičnih struktura i jezična svojstva unutar tekstova, dok se unutarnji kriteriji odnose na jezična svojstva tekstova, primjerice, formalnost. Kegalj i Borucinsky (2022) opisale su kriterije za sastavljanje specijaliziranoga usporednog i usporedivog korpusa (v. 1.5.) izvornih tekstova na engleskome jeziku i njihovih prijevoda na hrvatski jezik i izvornih hrvatskih tekstova, gdje je osnovni kriterij bio tematska domena pomorskoga prava. Pri tome su uvršteni sljedeći kriteriji:

Vanjski kriteriji:

1. *Cjelovitost* – korpusi se mogu sastaviti od cjelovitih tekstova ili tekstnih odsječaka; u ovome istraživanju prikupljeni su cjeloviti tekstovi zbog strukturne specifičnosti pravnih tekstova s jedne strane, te mogućnosti pojavljivanja nekih struktura samo u određenim dijelovima teksta (primjerice, u preambuli) s druge strane.
2. *Učinak* – ovaj kriterij se odnosi na cilj, funkciju ili učinak teksta u izvornom ili ciljnom okruženju. Cilj obuhvaća skupinu ljudi kojoj je tekst namijenjen, dakle, u prvome redu pomorcima i djelatnicima u pomorstvu. Prema funkciji tekstovi definiraju, objašnjavaju i provode pravila u specifičnom području ljudskoga djelovanja – pomorstvu, s čime je povezan i njihov pravni učinak, odnosno direktivnost pravnih propisa, koja bi trebala vrijediti za oba jezika (engleski i hrvatski).
3. *Žanr* – pod kriterijem žanra, uzeli su se u obzir tekstovi koji su institucionalni, tj. pravni tekstovi koji reguliraju područje pomorstva. Ovaj je kriterij povezan i s kriterijem izvora i pravnoga učinka, a obuhvaća i temu i vanjsku strukturu teksta, koja je specifična za svaki žanr, pa tako i pomorsko-pravni.
4. *Medij* – prema mediju, razlikuju se pisani, govoreni i elektronički tekstovi; u ovome su istraživanju u obzir uzeti isključivo pisani tekstovi.
5. *Raspon* – ovaj se kriterij odnosi na vremensko razdoblje u kojem su tekstovi nastali; s obzirom na to da je istraživanje sinkronijski usmjereno, u obzir su uzeti pravni tekstovi koji su na snazi.
6. *Izvor* – u ovome slučaju izvor su osobe, sastavljači pravnih tekstova i prevoditelji, odnosno institucije koje su donijele pravni propis. S obzirom na to da su i sastavljači i prevoditelji propisa razni i nepoznati, ta

odrednica nije bila relevantna, već se relevantnom smatrao status teksta kao institucionalnoga pravnog propisa.

7. *Status* – ovaj se kriterij odnosi na to je li tekst izvornik ili prijevod te je bio relevantan dvojako, prilikom odabira teksta za korpus, ali i prilikom uvrštavanja teksta u usporedni ili usporedivi korpus.

Unutarnji kriteriji:

1. *Tema* – kriterij teme, odnosno domene u koju tekst pripada, u ovome je slučaju bio prilično relevantan, jer se nisu uzimali u obzir tekstovi koji reguliraju druga pravna područja, već isključivo pomorsko pravo. Kao unutarnji kriterij, podrazumijeva internu analizu teksta, prvenstveno leksičkog aspekta kroz analizu ključnih riječi.
2. *Stil* – odnosi se na specifična strukturna ili leksička svojstva teksta i također proizlazi iz unutarnje analize teksta. U ovome istraživanju, u obzir se uzela formalnost tekstova, depersonaliziranost i jednosmjernost.

Na temelju gore definiranih kriterija, prikupljeni su tekstovi koji čine usporedni i usporedivi korpus pomorskopravnih tekstova, objedinjeni pod nazivom *MarLaw* (usp. Kegalj i Borucinsky, 2022). Korpus *MarLaw* čine usporedni korpus engleskih izvornika i njihovih prijevoda na hrvatski s ukupno 553 120 pojavnica za engleski i 498 389 pojavnica za hrvatski dio, te usporedivi korpus hrvatskih izvornih tekstova s ukupno 354 674 pojavnica. Korpus koji je sastavljen na ovaj način omogućuje kontrolirano promatranje jezičnih pojava jer ne dopušta interferencije žanrova i tema, a ipak omogućava promatranje jezičnih pojava iz više aspekata, supostavljajući izvornike i prijevode, isti žanr u dva jezika i prevedene i neprevedene tekstove iz kvantitativne i kvalitativne perspektive. Korpus *MarLaw* sastavljen je metodom ručnog prikupljanja tekstova (v. 4.3.), što omogućuje kontrolu kvalitete teksta, ali je dugotrajnije. Nadalje, ova vrsta specijaliziranoga korpusa zahtjeva i više-manje ručno sravnavanje tekstova u izvornom i ciljnom jeziku. Postupak sastavljanja višejezičnoga korpusa (v. 1.5.) nešto je zahtjevniji od sastavljanja jednojezičnoga korpusa[77] jer zahtijeva sravnavanje tekstova, koje je moguće provesti u računalnojezikoslovnim alatima kao što su *ParaConc* (Athelstan, 2010), *WordSmith* (Scott, 2008), *MemeSource, Hunalign*[78] i dr.

[77] Više o načinu sastavljanja paralelnih korpusa v. Build Parallel and Multilingual Corpora | Sketch Engine.

[78] *Memsource - Google Search, Hunalign – Sentence Aligner.*

4.4. IZRADA KORPUSA NA PRIMJERU ISTRAŽIVANJA ENGLESKIH RIJEČI U KORPUSU HRVATSKOGA JEZIKA

Kao što je prikazano u prethodnome poglavlju, prikupljanje tekstova i sastavljanje korpusa nije nimalo lak zadatak. Da sažmemo, prije sastavljanja korpusa bitno je odgovoriti na sljedeća pitanja[79]:

- Koji jezik, jezični varijetet ili funkcionalni stil uključiti?
- Gdje se mogu pronaći tekstovi?
- Kada su tekstovi objavljeni?
- Koje razdoblje će prikupljeni tekstovi obuhvaćati?
- Iz kojeg su područja tekstovi?
- Tko je napisao/objavio tekst?
- Kakve je jezične kvalitete tekst koji uključujemo u korpus?
- Koliko velik korpus treba biti?

U ovome se poglavlju na primjeru proučavanja engleskih riječi u hrvatskome prikazuje način izrade korpusa te poteškoće s kojima se istraživači susreću prilikom izrade i sastavljanja korpusa.

Iako se često tvrdi da je u europskim jezicima, pa tako i hrvatskome, sve više engleskih riječi, nije lako potvrditi ovu pretpostavku. Za početak valja definirati što smatramo engleskim riječima. U najširem smislu, ako prihvatimo Filipovićevu (1990) definiciju da se riječ posuđena iz engleskoga jezika koja označava neki pojam, ideju ili predmet vezan uz englesku civilizaciju naziva anglizmom, riječi koje istražujemo jesu anglizmi. Međutim, riječi posuđene iz engleskoga jezika jeziku primatelju, u ovome slučaju hrvatskome, prilagodile su se dijelom (npr. *bestseler* itd.) ili potpuno (ček, lizing itd.) (usp. također Duvnjak Jardas, 2019; Međeral, 2016), ili pak ostaju neprilagođene. Upravo se u novije vrijeme u hrvatskome sve češće javljaju engleske riječi koje zadržavaju izvorna grafijska obilježja, ali mogu primiti morfološka obilježja jezika primatelja, kao što je prikazano u primjeru (2).

> (2) *Prije nego li napravimo kratak* ***break****, pošaljite mi* ***linkove*** na objave, čekirajte zadnje ***mejlove*** *i* ***SMS****-ove i obavezno sejvajte* ***draftove*** kako vam se ne bi sve ***zbrejkalo*** dok ***ajtievci apdejtaju*** sustav.[80]

[79] Sastavljanje korpusa govornoga jezika zasebna je problematika. U ovoj se knjizi isključivo raspravlja o sastavljanju korpusa pisanoga jezika.

[80] Izvor: http://sretna.story.hr/kultura/link-na-najcesce-anglizme-u-hrvatskom-jeziku/.

Riječi koje su u primjeru (2) otisnute masnim slovima mogu se kategorizirati i kao pseudoanglizmi, odnosno riječi tvorene od elemenata preuzetih iz engleskoga jezika ili skraćenih engleskih riječi. Međutim, pojam pseudoanglizam često se poistovjećuje s lažnim anglizmima[81] pa se stoga u ovome istraživanju, naslanjajući se na Bogunović i Ćoso (2013); Brdar (2010); Ćoso i Bogunović (2017) preferira naziv engleske riječi, a odnosi se na riječi koje su preuzete iz engleskoga jezika s izvornim grafijskim obilježjima, a mogu se pojaviti s hrvatskim morfološkim nastavcima. Budući da je posuđivanje engleskih riječi prije svega prisutno na leksičkoj razini, ovim istraživanjem[82] nisu obuhvaćene gramatičke riječi (npr. *in, out* itd.). Nadalje, engleskim riječima ne smatramo ni sljedeće kategorije:

1. riječi iz stranoga jezika koji nije engleski (npr. *Welt* iz njemačkoga, *pour* iz francuskoga)
2. vlastita imena ili toponimi (npr. *Wave Boat Sealver, Cambridge*)
3. lažni anglizmi (npr. *gastro show*)
4. pogrešno označene riječi (hrvatski prijedlog *van* označen kao engleska imenica *van* 'kamion')
5. kratice i akronimi (npr. *btw, USA*)
6. pogrešno napisane riječi (npr. *coffe*)
7. idiosinkratične pojave (npr. *pussyhit*).

Hipotezu „Engleske se riječi sve češće javljaju u hrvatskome" mogli bismo operacionalizirati na nekoliko načina. Kao prvo, mogli bismo intuitivno popisati engleske riječi koje smatramo da se pojavljuju u hrvatskome. Ova metoda je introspektivna i subjektivna, i postoji velika vjerojatnost da se ne bismo ni približno sjetili većine engleskih riječi koje se pojavljuju u hrvatskome. Drugi način mogao bi obuhvatiti prikupljanje odgovora ispitanika o tome što misle koje su engleske riječi prisutne u hrvatskome pri čemu bi, između ostaloga, trebalo odrediti veličinu uzorka ispitanika s obzirom na broj govornika hrvatskoga jezika te

[81] Pseudoanglizmi, sekundarni ili prividni anglizmi riječi su unutar hrvatskoga jezika tvorene od engleskih elemenata ili od engleskih riječi skraćenih u novi lik (Filipović, 1990), pri čemu treba naglasiti da kao takve nisu preuzete iz engleskoga već se slobodno oblikuju u jeziku primaocu (Muhvić-Dimanovski i dr. 2005) – u popisu literature samo Muhvić-Dimanovski 2005?. Također se nazivaju i lažnim anglizmima i kao takvi ne postoje u jeziku davatelju, a primjere nalazimo u mnogim jezicima. U njemačkome se, primjerice, rabi imenica *Handy* (engl. *mobile phone* 'mobitel'), u nizozemskome *beamer* (engl. *video projector* 'projektor'), u francuskome *pompon girl* (engl. *female cheerleader* 'navijačica'), *le baby-foot* (engl. *table-football* 'stolni nogomet'), *des baskets* (engl. *sneakers, trainers, tennis shoes* 'tenisice'), u hrvatskome smoking (engl. *tuxedo*), u hrvatskome i poljskome dres (engl. *track suit*) itd.

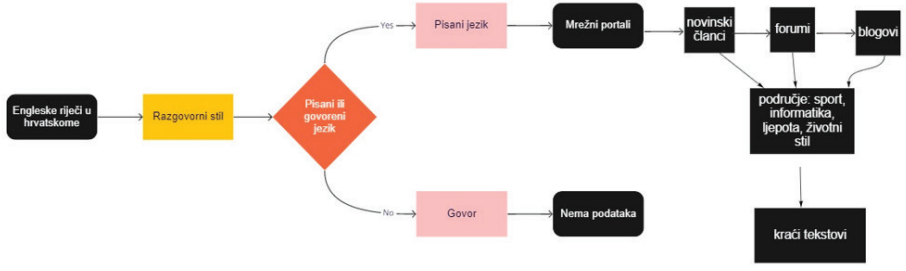
[82] O načinu crpljenja engleskih riječi iz korpusa hrvatskoga jezika raspravlja se u poglavlju 6.

demografski sastav s obzirom na to da su engleske riječi novija pojava i pretpostavlja se da ih mlađe generacije više koriste u svakodnevnoj komunikaciji. Treći način bio bi da ručno pregledavamo tiskovine i bilježimo engleske riječi koje su pojavljuju u hrvatskim tekstovima, što bi bilo vrlo dugotrajno. Ovim se načinom, primjerice, pratio priljev stranih i engleskih riječi u hrvatskome i poslužio je pri izradi rječnika neologizama (Muhvić-Dimanovski, Skelin Horvat i Hriberski, 2016). Četvrti se način odnosi na pregledavanje postojećih popisa engleskih riječi u rječnicima, na wikipediji, raznim portalima i blogovima. Ova bi metoda bila također vrlo dugotrajna i nesustavna jer riječi nisu na temelju istih kriterija ušle na popise, a većina popisa obuhvaća popis anglizama (npr. *wikipedija*.*hr*) što znači da na njemu ne bismo pronašli (neprilagođene) engleske riječi koje nas zanimaju. Peti je način da na temelju jezičnoga uzorka, tj. korpusa, pokušamo potvrditi ili opovrgnuti danu hipotezu. U nastavku se opisuje potonji način, tj. pokušaj da se na temelju korpusa odredi koje se engleske riječi javljaju u hrvatskome jeziku. Potom valja odabrati jezični uzorak ili korpus na kojemu ćemo ispitati ovu jezičnu pojavu. Idealno bi bilo sastaviti jezični uzorak govornih tekstova, no financijski je neisplativo, a i sa stajališta ljudskih resursa također problematično. Stoga odabiremo jednostavnije rješenje, a to je korpus pisanoga jezika. Za hrvatski jezik preko računalnojezikoslovnih alata *SkE* i *NoSkE* dostupni su korpusi *hrWaC, HNK* i *Riznica*. *Riznica* ne bi bio dobar jezični uzorak jer obuhvaća pisane tekstove od 11. stoljeća do danas. *HNK* je zastario, a *hrWaC* je prikupljan od 2011-2013, dakle, pred više od 10 godina te nije reprezentativan u pravome smislu riječi (v.3). Stoga smo odlučili sastaviti vlastiti korpus za potrebe ovoga istraživanja. Znanstveni radovi koji se bave tematikom anglizama i engleskih riječi pokazuju da se takve riječi najčešće javljaju u medijima (npr. Alvarez-Mellado, 2020; Brdar, 2010; Runjić-Stoilova i Pandža, 2010 i dr.), što znači da bismo mogli sastaviti korpus koji će predstavljati novinski funkcionalni stil. U nastavku su prikazane odluke koje su donesene prilikom sastavljanja i izrade korpusa:

- ✓ Koji jezik, jezični varijetet ili funkcionalni stil uključiti? – novinski funkcionalni stil, kraći tekstovi
- ✓ Gdje se mogu pronaći tekstovi? – na mrežnim portalima, blogovima, forumima
- ✓ Kada su tekstovi objavljeni? – od 2005. nadalje
- ✓ Koje razdoblje će prikupljeni tekstovi obuhvaćati? 2015. – 2022.
- ✓ Iz kojeg su područja tekstovi? – sport, informatika, ljepota, životni stil itd.
- ✓ Tko je napisao/objavio tekst? – novinari, nepoznato, tj. pojedinci iz opće populacije

- ✓ Kakve je jezične kvalitete tekst koji uključujemo u korpus? – nepoznato, postoji mogućnost da se u korpusu pojave i strojno prevedeni tekstovi, što će ovisiti o načinu izrade korpusa (tj. povlačenjem URL-a ili zadavanjem definiranih riječi na temelju koji će *WebBootCaT* tehnologija prikupiti tekstove, v. 4.3.)
- ✓ Koliko velik korpus treba biti? – nekoliko milijuna riječi

Odluke i odabiri koji su učinjeni prilikom sastavljanja korpusa koji smo nazvali *Engleske riječi u hrvatskome* (*ERH*) prikazane su na slici 12.



Slika 12: *Prikaz faza u izradi korpusa* ERH

Sljedeća se odluka tiče načina na koji će korpus biti sastavljen. Odlučeno je da će se korpus sastaviti automatski jer je to jednostavnija i brža metoda. Jedan od načina automatskoga sastavljanja jest da se unesu početne riječi (engl. *seed words*) i da su definirane riječi[83] unutar domene .hr. Budući da ne

[83] *Seed words:* blog (internetski dnevnik, digitalni dnevnik, e-dnevnik), bookmark (oznaka mjesta čitanja, straničnik), brand (robna marka), break (stanka, pauza), bullying (nasilje, nasilje u školi, vršnjačko nasilje), chat (virtualno/internetsko brbljanje, čavrljanje, ugodan razgovor, e-razgovor), celebrity (poznata, slavna osoba, zvijezda), cruiser (krstaš, brod koji krstari), display (predočnik), e-mail (e-pošta, elektronička pošta), event (događaj, događanje), fashion week (modni tjedan), last minute (u posljednji trenutak), leasing (iznajmljivanje, ustupanje, davanje u najam, zakup), link (poveznica), mobbing (zlostavljanje), monitoring (motrenje, promatranje, nadzor, praćenje), mystery shopping, mystery shopper (tajno kupovanje, tajni kupac), nick (nadimak), party (zabava, proslava, domjenak), shopping (kupovina, kupnja), shopping centar (trgovački centar), stage (pozornica), tender (natječaj, ponuda), website (internetska stranica), accessoire/asesoar (modni dodatak/detalj/pribor), all-inclusive (sveobuhvatan/sa sveobuhvatnom ponudom), backup (sigurnosna kopija), benefit (povlastica, prednost, korist), cost-benefit analiza (analiza troška i koristi, istaknuo je prednosti (benefite) svojega programa), big data (veliki podaci), brain-drain (odljev mozgova), case study (analiza/studija slučaja), crumble (mrvljenac), factoring (otkup nedospjelih potraživanja), gadget ((pametna) spravica), happy hour (vrijeme nižih cijena, vrijeme sniženja), host - hostesa (domaćica i domaćin), life coach (životni trener), mainstream mediji (mediji glavne struje, prevladavajući mediji), teaser (mamac), widget (mala aplikacija). Izvor: Interpreta.hr.

postoji popis engleskih riječi u hrvatskome, mogli bismo potražiti u korpusu engleskoga najčešće engleske riječi, no to ujedno neće biti i najčešće engleske riječi u hrvatskome. Za početak smo, stoga, s bloga *interpreta.hr* preuzeli popis najčešćih anglizama u hrvatskome koje smo unijeli kao početne riječi. Ovaj popis najbliže odgovara pojmu engleskih riječi kako su definirane ovim istraživanjem. Popis, međutim, nije određen temom, ali prema subjektivnoj procjeni pokazuje engleske riječi koje su u svakodnevnoj uporabi u hrvatskome pa ovaj način izrade korpusa predstavlja tek prvi pokušaj.

Unutar 60-ak minuta *SkE* je pomoću tehnologije *WebBootCaT* (v. 4.2.) prikupio i označio korpus koji smo nazvali *Engleske riječi u hrvatskome*_01 (*ERH* v1[84]) koji sadrži 13 821 954 pojavnica. Iako taj broj zvuči vrlo impresivno, valja provjeriti kvalitetu takva korpusa pomoću opcije crpljenja terminologije (v. 3.), koja funkcionira na način da se iz fokusnoga ili zadanoga korpusa (engl. *focus corpus*) riječi uspoređuju s referentnim korpusom (engl. *reference* corpus, u ovome slučaju *hrWaC*-om,), a služi kao prva provjera kvalitete sastavljenoga korpusa. Valja imati na umu da prilikom prvog sastavljanja korpusa nećemo dobiti optimalne rezultate, te da će biti potrebno ponoviti postupak kompilacije korpusa. Tako iz slike 13 možemo iščitati da među ključnim riječima u sastavljenome korpusu ima pogrešno napisanih riječi (*lanak* umjesto članak) ili riječi s dijakritičkim znakovima koje označivač nije prepoznao, riječi iz stranoga jezika koji nije engleski (*zlaty*), pogrešno označenih riječi te cijelih rečenica na engleskome, iako smo nastojali sastaviti korpus hrvatskoga jezika s engleskim riječima ili frazama, a ne cijelim rečenicama ili tekstovima na engleskome. Načelno, računalu je vrlo jednostavno odrediti kojem jeziku neka riječ, fraza ili rečenica pripada, u što se možemo uvjeriti upišemo li bilo koju frekventnu hrvatsku riječ u programu *Google Translate* (*GT*)[85], ali je teško pronaći zasebne elemente stranoga (u ovome slučaju engleskoga) jezika (engl. *foreign/English inclusions*) u tekstu (usp. Alex, 2005).

[84] Korpus je sastavljen 10. studenoga 2021.

[85] Izvor: Google Translate (GT). O načinu na koji GT detektira i prepoznaje riječi nekog jezika mnogo se raspravljalo, a raspravlja se još i danas. Pretpostavlja se da je prepoznavanje temeljeno na nadziranome sustavu u kombinaciji s heurističkim modelima koji proučavaju kako je tekst kodiran. O razvoju algoritama za identifikaciju jezika v. primjerice Jauhiainen i dr., 2019. Naravno, koliko god napredne ove tehnologije bile, nisu potpuno točne. Pa tako, ako primjerice u GT upišemo riječ slon, prepoznat će je kao riječ iz češkoga jezika, riječ orao kao riječ iz bosanskoga jezika, a riječ naputak kao riječ iz hrvatskoga jezika.

Slika 13: *Ključne riječi iz ad hoc sastavljenoga korpusa* ERH v.1

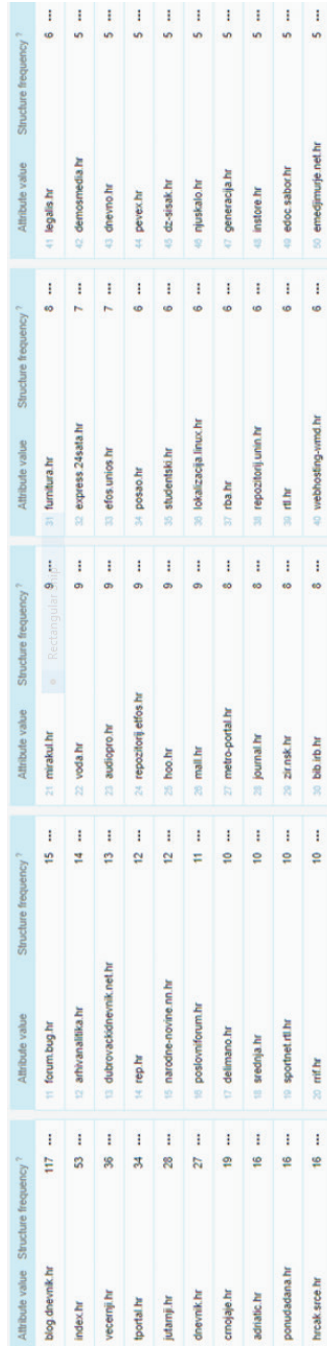
Prva je riječ na popisu ključnih riječi slovenska riječ *kar*, a uz nju, kao što je razvidno iz slike 13, ima puno neželjenoga šuma (kratice *Oct*, *wed* koje su dijelom nejezičnoga sadržaja, interpunkcijski znakovi kao ključne riječi), dupli podaci i sl.

Nadalje, iako smo prilikom kompilacije korpusa definirali da se tekstovi povlače s domene .hr, i očekivali da ćemo dobiti tekstove na hrvatskome, s obzirom na to da su mnoge hrvatske (službene) mrežne stranice prevedene na engleski, u korpus su uključene i cijele rečenice, odlomci i tekstovi na engleskome, što znatno može utjecati na rezultate korpusne pretrage.

Idući problem koji se javlja jest dvostruki sadržaj. Iz tog je razloga korpus ponovno kompiliran, uz opciju da ukloni sve duplikate.

Iz slike 14 razvidno je da je korpus sastavljen i od mrežnih stranica na kojima ne očekujemo velik broj engleskih riječi u hrvatskome, kao što su primjerice *narodne-novine.nn.hr*, *edoc.sabor.hr*, *sskranjcevic.hr* i sl. Stoga su takve stranice ručno uklonjene.

Slika 14: *Popis web stranica koje čine korpus* ERH v1



	Attribute value	Structure frequency
1	blog.dnevnik.hr	117
2	index.hr	53
3	vecernji.hr	36
4	tportal.hr	34
5	jutarnji.hr	28
6	dnevnik.hr	27
7	cimgajge.hr	19
8	adriatic.hr	16
9	ponudadana.hr	16
10	hrcak.srce.hr	16
11	forum.bug.hr	15
12	arhivanalitika.hr	14
13	dubrovackidnevnik.net.hr	13
14	rep.hr	12
15	narodne-novine.nn.hr	12
16	poslovniforum.hr	11
17	delimano.hr	10
18	brednja.hr	10
19	sportnet.rtl.hr	10
20	mf.hr	10
21	mirakul.hr	9
22	voda.hr	9
23	audiopro.hr	9
24	repozitorij.etfos.hr	9
25	hoo.hr	9
26	mali.hr	9
27	metro-portal.hr	8
28	journal.hr	8
29	zsr.nsk.hr	8
30	bib.irb.hr	8
31	furnitura.hr	8
32	express.24sata.hr	7
33	efos.unios.hr	7
34	posao.hr	6
35	studentski.hr	6
36	lokalizacija.linux.hr	6
37	roa.hr	6
38	repozitorij.unin.hr	6
39	rtl.hr	6
40	webhosting-vimd.hr	6
41	legalis.hr	6
42	demosmedia.hr	5
43	dnevno.hr	5
44	pevex.hr	5
45	dz-sisak.hr	5
46	rijukalio.hr	5
47	generacija.hr	5
48	instore.hr	5
49	edoc.sabor.hr	5
50	emedjimurje.net.hr	5

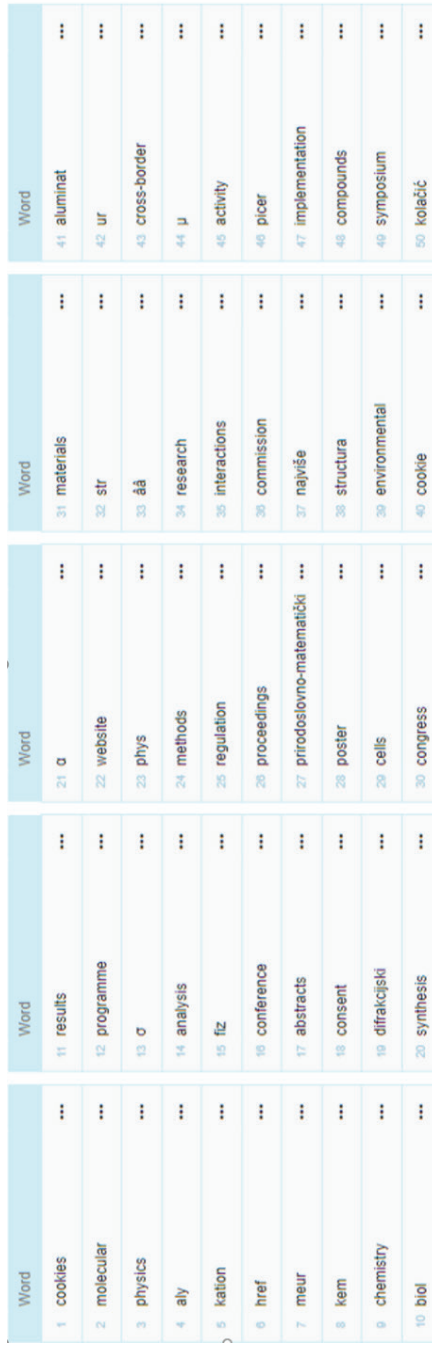
Iz navedenoga slijedi da sastavljeni korpus *ERH v1* ne možemo smatrati odgovarajućim jezičnim uzorkom za jezičnu pojavu koju želimo ispitati te da je potrebno ponovno sastaviti korpus uz podešene postavke. Za početak, potrebno je bolje definirati riječi na temelju kojih će tehnologija birati tekstove. Također, *SkE* nudi nekoliko opcija koje mogu biti korisne da bi se izbjegle višeznačnosti u pretrazi. Primjerice, korisnik može odrediti da određene riječi budu obavezne, tj. da moraju biti prisutne u dokumentu koji se traži na mreži, tzv. funkcionalnost *allowlist*, dok istovremeno iz pretrage može isključiti određene riječi, tzv. funkcionalnost *denylist*.

Iz razloga što prva inačica korpusa nije dala zadovoljavajuće rezultate, korpus smo rekompilirali ili ponovno sastavili koristeći sljedeće postavke:

- ukloni duplikate
- od 1 321 mrežnih stranica i dokumenta s kojih su crpljeni tekstovi, izbrisano je 249 (zakoni, stranice na engleskome)
- izbaci hrvatske nazive iz popisa riječi pod fusnotom 83.

Druga inačica korpusa *ERH* ima znatno manji broj pojavnica (2 659 274) u odnosu na prvu inačicu, ali kao što je razvidno iz slike 15 i opcije crpljenja riječi, ni ova inačica korpusa nije zadovoljavajuća te još uvijek nisu razriješeni problemi duplikata i neadekvatnih izvora.

Slika 15: *Ključne riječi korpusa* ERH v2



Sljedeći način rješavanja ovoga problema sastoji se u tome da se točno definiraju mrežne stranice s kojih želimo povlačiti tekstove i razdoblje u kojem su tekstovi nastali (v. 4.2., usp. također Kučić, 2021).

Za potrebe ovoga istraživanja, i prikaza sastavljanja korpusa na način da se povlači URL odabrana jer mrežna stranica *Hrportali* koja sadrži popis svih portala u Hrvatskoj.

Iz navedenoga možemo zaključiti da je *SkE* vrlo moćan alat, ali da je pomno planiranje i promišljanje potrebno kako bi se sastavio prihvatljiv korpus. Sva tri načina sastavljanja korpusa, odnosno ručno sastavljanje i podizanje vlastitog teksta u *SkE*, te automatsko sastavljanje preko URL-a ili mrežnih stranica i definiranjem riječi moguće je kombinirati, no valja svakako imati na umu nedostatke ovakva korpusa, kao što su reduplikacija, neželjeni šum, promjena sadržaja na internetu, što u konačnici može utjecati na pouzdanost i mjerljivost dobivenih rezultata.

U nastavku ćemo prikazati i kritički se osvrnuti na vrste podataka koji se mogu dobiti iz korpusa, grupirajući ih u tri cjeline: evidencija, frekvencija i relacija, te na suodnos leksika i gramatike kroz korpusnolingvistički pogled.

5. EVIDENCIJA I FREKVENCIJA

Evidencija je popis jezičnih jedinica koje možemo dobiti iz korpusa, a pokazuje nam nalaze li se X i Y u korpusu. Tako možemo provjeriti nalazi li se određena jezična jedinica (tj. određena riječ ili imenica, glagol, pridjev, prilog, sintagma, kolokacija, frazem i dr.) u uzorku koji smo sastavili za potrebe vlastitoga istraživanja ili u uzorku koji je već dostupan. Ovakve pretrage u pravilu su najjednostavnije jer ćemo u tražilicu upisati traženu riječ i provjeriti pojavljuje li se u korpusu. Na taj način možemo tražiti i strane riječi, npr. *freelancer*, *e-mail* itd., odnosno sve riječi za koje znamo ili pretpostavljamo da će se naći u danome uzorku. Većina računalnojezikoslovnih alata, pa čak i oni jednostavniji, mogu prikazati traženu riječ s kontekstom u kojemu se pojavljuje, odnosno konkordancije (v. poglavlje 3). Kontekst tražene riječi omogućuje daljnju jezičnu pretragu i u pravilu kvalitativnu analizu na način da pregledamo konkordancije riječi, sve ili samo nasumične, što će ovisiti o broju dobivenih rezultata te količini konkordancijskih nizova ili rečenica koje istraživač može pregledati.

Iz tablice 5 možemo iščitati da se riječi *ključić* i *prijestolonasljednik* nalaze u tri korpusa hrvatskoga jezika, odnosno u korpusu standardnoga jezika (*Riznica*), korpusu općega jezika (*hrWaC*) te u korpusu hrvatskih internetskih portala (*ENGRI*), dok su sintagme *računalni oblak* i *USB stick* prisutne samo u dva korpusa (*ENGRI* i *hrWaC*), a fraza *strašno temeljit* samo u jednom korpusu (*ENGRI*). Međutim, na temelju rezultata iz tablice 7 ne možemo tvrditi da fraze *računalni oblak*, *USB stick* i *strašno temeljit* nisu dijelom hrvatskoga jezika ili da ih govornici hrvatskoga jezika ne rabe. Ako korpus predstavlja samo jezični uzorak (engl. *sample*) a ne jezik u cjelini (engl. *population*[86]), u tom slučaju odsustvo evidencije u korpusu nije dokaz o nepostojanju takve riječi ili fraze.

[86] U korpusnim istraživanjima u pravilu proučavamo uzorak, a ne cijeli jezik jer je nemoguće proučavati jezik u cjelini. O cjelini (ili populaciji u statističkome smislu) možemo govoriti ako, primjerice, proučavamo djela M. Marulića te ako smo sastavili korpus od svih tekstova koje je pisac napisao.

Tablica 7: *Evidencija (i brojanje) riječi u općima korpusima hrvatskoga jezika*

Riječ ili fraza / Korpus	**Riznica (100 milijuna riječi)**	**ENGRI (900 milijuna riječi)**	**hrWaC (1,9 milijardi riječi)**
računalni oblak	0	28	103
prijestolonasljednik	100	1677	745
ključić	85	177	741
USB stick	0	312	2224
strašno temeljit	0	1	0

Kod evidencije u korpusu tražimo potvrde za postavljene hipoteze, npr. u kojem se kontekstu i kojim tekstovima rabi sintagma *računalni oblak*, koriste li se u hrvatskome riječi *celebrity, zatipak* ili *tipfeler*? Pomislili bismo da bi slanje jednostavnoga upita *Google* tražilici moglo dati odgovor na pitanje preferiraju li govornici hrvatskoga jezika riječ *tipfeler* ili *zatipak.* Za prvu je dobiveno 51 000 rezultata, a za potonju 3 440, iz čega bismo mogli zaključiti da govornici hrvatskoga jezika preferiraju riječ *tipfeler* u odnosu na novotvorenicu *zatipak*, što u pitanje dovodi potrebu za razvojem računalnojezikoslovnih alata. Međutim, valja imati na umu da *Googleovi* algoritmi nisu isti za svaki dio svijeta, svakoga čovjeka i svaku pretragu, i razlikuju se npr. i prema tome je li korisnik prijavljen ili nije. Ne možemo dakle brojeve dobivene pretragom *Googlea* uzeti kao apsolutno točne, a glavni razlog je taj da ne znamo na koji način je *Google* dobio, pa tako i korisniku prikazao, rezultat, odnosno ne kontroliramo niti što je pretraženo niti iz kojeg izvora, dok ove parametre u korpusu možemo kontrolirati.

Nadalje, vrijednost je računalnojezikoslovnih alata u evidenciji, frekvenciji, relaciji i drugim vrstama podataka koje možemo dobiti iz korpusa ili iščitati iz konkordancija, lista riječi, skica riječi, tezaurusa i sl. (v. 3) te mogućnost kontrole onoga što se pretražuje i kako se pretražuje, što će biti detaljno prikazano u nastavku.

Primjerice, ako nas zanima koristi li se u hrvatskome posuđenica *burnout* (također *burn out*) ili njezina odgovarajuća zamjena u hrvatskome (*izgorjelost, izgaranje*), do tog ćemo podatka lako doći u korpusu, kao i do mogućih kolokacija (npr. *sindrom izgorjelosti, sindrom burnout(a)*). Problem nastaje kada želimo pronaći sve ili po mogućnosti većinu stranih odnosno engleskih riječi u korpusu hrvatskoga jezika, a ne znamo točno koje bi to riječi mogle biti. Mogli

bismo ih intuitivno popisati ili pretraživati rječnike neologizama (pod uvjetom da se redovito ažuriraju) ili pomoću upitnika saznati koje engleske riječi govornici hrvatskoga najčešće koriste u svakodnevnoj komunikaciji te ih pojedinačno pretraživati u korpusu. Očito je da bi takvo što funkcioniralo za stotinjak (najčešćih) engleskih riječi, no bilo što iznad toga bilo bi vremenski neisplativo. Nadalje, ovom metodom bismo lako mogli propustiti neke engleske riječi koje se nalaze u korpusu hrvatskoga jezika, a do kojih nismo došli pukom intuicijom.[87]

Tražimo li evidenciju riječi u korpusu, moramo uzeti u obzir da će, primjerice, jednostavna pretraga (engl. s*imple search*) riječi *ugrađen* dati 2437 rezultata, a *ugrađeni* 1806 rezultata, pri čemu će *ugrađen* biti dijelom glagola ili imenice, a zbog sinkretizma oblika (množina i određeni pridjev) valja jasno postaviti korpusnu pretragu. Sustav će tako izlistati konkordancije samo za pojave u nominativu (npr. *ugrađeni novi sustav, ugrađeni novi mehanizam*), a ako želimo dobiti rezultate i u kosim padežima bit će potrebno definirati sljedeću pretragu: [word="ugrađen|ugrađeno|ugrađena|ugrađeni|ugrađena|ugrađenih"] itd., dakle potrebno je navesti sve moguće padežne nastavke. Drugi način na koji možemo doći do rezultata jest da pretragu definiramo na sljedeći način: prikaži lemu *ugraditi*, a da nije označena kao glagol ili prikaži lemu *ugraditi* a da je označena kao pridjev, tj. [lemma="ugraditi" & !tag="V.*"] ili [lemma="ugraditi" & tag="A.*"], što neće dati iste rezultate ni apsolutno točan rezultat, no može biti dobar pokazatelj. Nadalje, tražimo li u korpusu evidenciju ili frekvenciju fraze *najbolja ponuda*, pretragu je potrebno sastaviti na sljedeći način: [lemma="dobar" & tag="Ags.*"] [lemma="ponuda"]. Također, za pojedine imenice lema je nešto drugačija od one koju bismo očekivali. Tako je, primjerice, lema imenice *bicikl* u korpusima *ENGRI, hrWaC* i *Riznica bicik*, a ne *bicikl,* a lemu ćemo pronaći na način da u KWIC formatu odaberemo prikaz lema ispod pojavnice. To je važno znati, i tokom istraživanja važno je rješavati ovakve probleme na koje se naiđe u korpusu. Navedeni primjeri ilustracija su rada na *Bazi engleskih riječi i hrvatskih istovrijednica* (Bogunović, Jelčić Čolakovac i Borucinsky, 2022).

Frekvencija je, uz evidenciju, vrsta osnovnih podataka koji se mogu dobiti iz korpusa, a pokazuje koliko puta se X pojavljuje u korpusu. Frekvencije predstavljaju polazište za istraživanja vokabulara, poučavanje jezika, leksikološka, leksikografska, terminološka i druga istraživanja, i valjan su razlog za

[87] Opreka između intucije i empirije u korpusnoj lingvistici često se označava kao opreka između tzv. *armchair linguists* i *corpus linguists* (Fillmore, 1992).

investiranje u razvoj jezičnih tehnologija. Analiza teksta kao zbirke podataka te mogućnost dobivanja statističkih podataka omogućuje kompleksniju analizu i pretragu jezika nego što bi to bilo moguće nekorpusnim metodama. Upravo je u tome odgovor na pitanje zašto investirati u računalnojezikoslovne alate i koje su prednosti korpusnih metoda u jezikoslovnim istraživanjima.

Nadalje, frekvencije mogu pokazati tendencije koje postoje u jeziku, u prijevodima i sl. Jednostavnom pretragom funkcionalnosti lista riječi možemo utvrditi da su najčešće leme u *hrWaC-u biti*, *i*, *u*, *sebe*, *da*, najčešće POS oznake *Z*, *RGP*, *CC*, *Var3s*, *Sl*[88], najčešće imenice *godina*, *čovjek*, *dan*, *vrijeme*, *Hrvatska*, najčešći glagoli *biti*, *htjeti*, *moći*, *imati*, *trebati*, a pridjevi *sav*, *velik*, *nov*, *dobar*, *hrvatski*, itd.

Tražimo li najčešće imenice u korpusu, jednostavna statistička mjera koju u *SkE*-u pronalazimo pomoću opcije *lista riječi* (v. 3) dat će tražene podatke. Tako iz dobivenih rezultata možemo iščitati da su najčešće imenice u korpusu hrvatskoga općeg jezika (*hrWaC*, slika 16) i standardnoga jezika (*Riznica*, slika 17) *godina*, *čovjek*, *dan*, *vrijeme* itd.

[88] Oznaka Z stoji za interpunkcijski znak, Rgp je opći pridjev u pozitivu, Cc je koordinacijski veznik, Var3s pomoćni glagol u 3. licu jednine, Si prijedlog iza kojeg slijedi instrumentalni oblik.

Croatian Web (hrWaC 2.2, ReLDI)

358,984,147 total frequency)

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	godina	4,075,743	11	problem	1,073,711	21	broj	837,991	31	žena	673,501	41	klub	575,794
2	čovjek	2,310,144	12	rad	1,065,635	22	sat	819,629	32	mjesec	668,902	42	zakon	574,323
3	dan	2,117,622	13	kraj	948,562	23	riječ	790,410	33	film	665,630	43	milijun	570,549
4	vrijeme	1,726,649	14	način	937,988	24	strana	771,305	34	predsjednik	657,092	44	utakmica	566,315
5	hrvatska	1,573,843	15	pitanje	891,593	25	zagreb	761,106	35	područje	621,191	45	član	565,329
6	dio	1,259,387	16	svijet	882,170	26	pravo	719,506	36	projekt	620,189	46	društvo	560,667
7	mjesto	1,190,743	17	zemlja	875,304	27	slučaj	709,756	37	škola	616,961	47	rezultat	549,558
8	dijete	1,176,425	18	posao	866,492	28	kuna	701,824	38	cijena	604,051	48	ime	535,885
9	život	1,157,916	19	stvar	865,524	29	država	697,925	39	kuća	599,662	49	vlada	534,013
10	grad	1,131,303	20	osoba	853,858	30	program	693,807	40	sustav	597,968	50	prostor	528,051

Slika 16: *Najfrekventnije imenice u korpusu* hrWaC

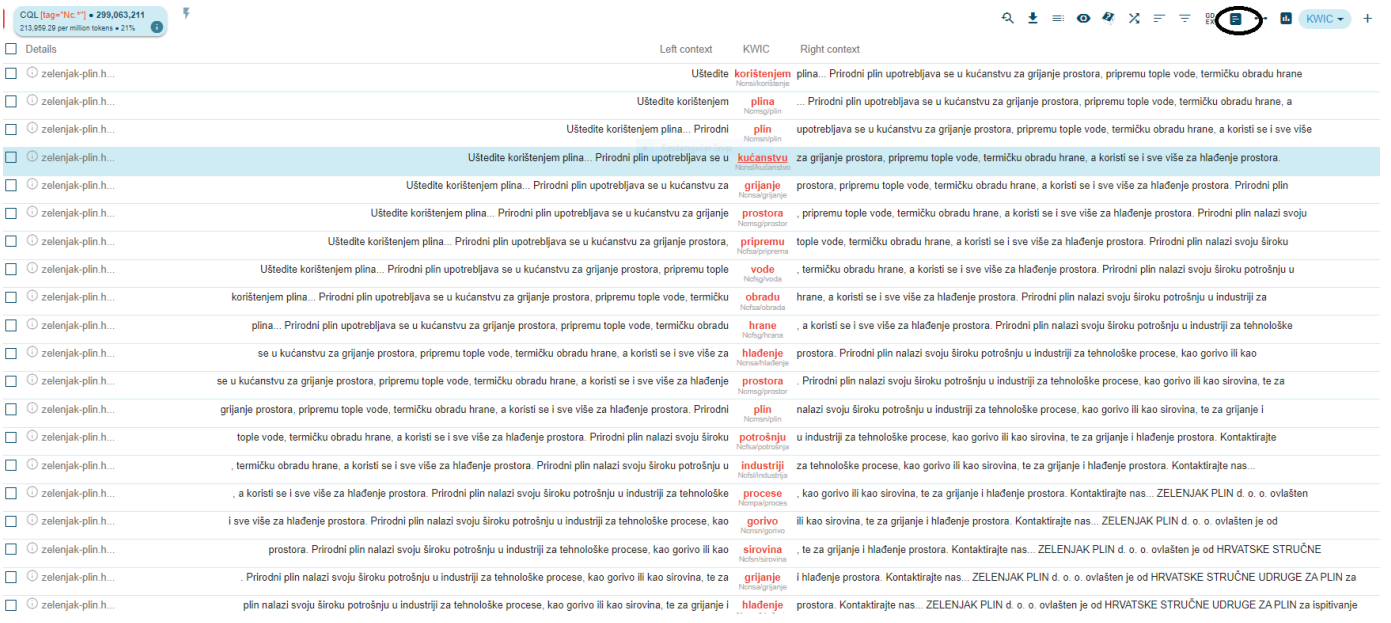
Riznica v0.1

27,899,095 total frequency)

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	godina	307,268	11	dio	92,527	21	posao	68,325	31	način	55,213	41	kuća	50,169
2	hrvatska	211,864	12	mjesto	91,904	22	utakmica	65,326	32	stranka	55,080	42	vlast	48,774
3	zagreb	167,645	13	kuna	84,128	23	problem	64,689	33	osoba	54,408	43	mjesec	48,394
4	dan	138,599	14	zakon	78,884	24	pravo	63,936	34	dijete	54,127	44	odluka	47,942
5	čovjek	113,764	15	rad	77,185	25	slučaj	63,411	35	sat	52,183	45	igrač	47,859
6	predsjednik	111,679	16	grad	76,293	26	kraj	63,236	36	članak	52,086	46	rat	47,112
7	zemlja	101,368	17	država	75,094	27	broj	60,219	37	ministarstvo	51,663	47	program	46,378
8	vrijeme	100,942	18	milijun	74,832	28	ministar	59,193	38	odnos	51,401	48	područje	46,338
9	vlada	98,721	19	pitanje	73,426	29	svijet	57,881	39	strana	51,346	49	rezultat	45,055
10	riječ	92,616	20	sud	68,378	30	život	55,480	40	izbor	50,214	50	član	45,019

Slika 17: *Najfrekventnije imenice u korpusu* Riznica

Na isti način možemo dobiti najčešće pridjeve, priloge, glagole i druge vrste riječi. Kao što je razvidno iz slike 18, dobiveni popisi osim općih imenica sadrže i vlastite (npr. *Hrvatska*), a želimo li dobiti najčešće opće imenice, pretragu možemo provesti i preko opcije *konkordancija* te je suziti na način da definiramo samo opće imenice (engl. *common noun*), odnosno upitom [tag="Nc.*"]. Ova pretraga u korpusu *hrWaC* polučila je više od 306 milijuna rezultata, a alat prikazuje samo nasumično odabranih 10 milijuna konkordancijskih nizova na temelju kojih se računa frekvencija, pa stoga rezultati pretrage frekvencija preko konkordancija i liste riječi neće biti identični (usp. sliku 18).



Slika 18: *Konkordancijski niz pretrage općih imenica* [tag="Nc.*"] u hrWaC-u

Show relative frequency ☐ Show percentage of concordance lines

(184,568 items, 10,000,000 total frequency)

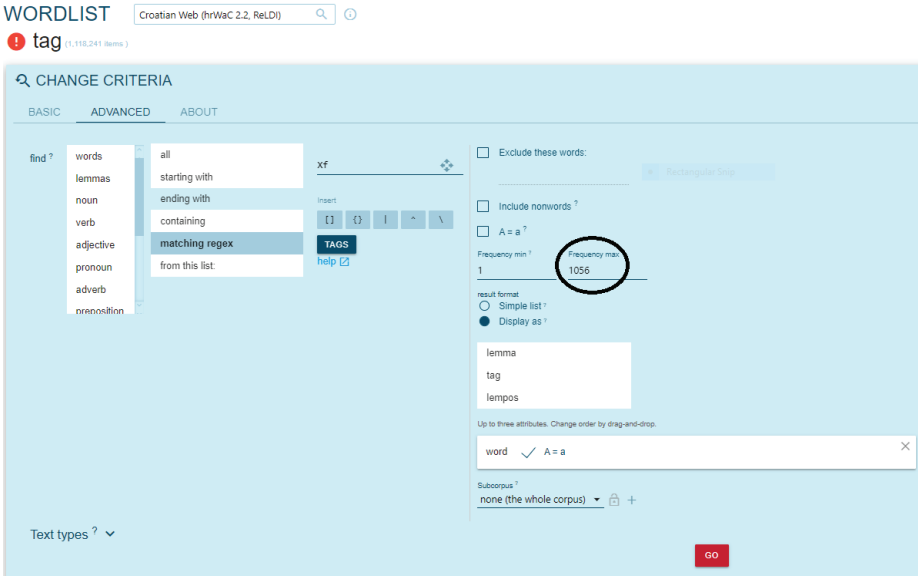
		Lemma	Frequency ↓	Relative ?
1	☐	godina	145,448	103.46
2	☐	dan	67,404	47.95
3	☐	čovjek	55,676	39.60
4	☐	vrijeme	53,723	38.22
5	☐	rad	53,049	37.74
6	☐	dijete	45,270	32.20
7	☐	dio	42,351	30.13
8	☐	život	41,103	29.24
9	☐	mjesto	40,967	29.14
10	☐	grad	39,731	28.26
11	☐	osoba	35,622	25.34
12	☐	način	35,450	25.22
13	☐	program	35,001	24.90
14	☐	područje	31,791	22.61
15	☐	broj	30,739	21.87
16	☐	sat	30,413	21.63
17	☐	projekt	30,238	21.51
18	☐	svijet	29,277	20.83
19	☐	škola	28,153	20.03
20	☐	sustav	27,753	19.74

You are only allowed to access 1,000 items. Get more

Rows per page: 20 1–20 of 1,000 1 / 50

Slika 19: *Frekvencija općih imenica u korpusu* hrWaC

SkE ima ograničenje na preuzimanje samo prvih 1000 najčešćih riječi, no tome se može doskočiti na način da se podese maksimalna i minimalna frekvencija, tj. upiše apsolutna frekvencija zadnje riječi s popisa koji smo dobili kao što je prikazano na slici 19 pa se rezultati ponovno preuzmu i taj postupak ponovi dok ne dobijemo sve željene frekvencije. Ovdje je to prikazano na primjeru oznake *Xf* u korpusu *hrWaC*. Prvih 1000 rezultata smo spremili, zatim pod maksimalnu frekvenciju upisali posljednji broj iz tablice, 1056, u idućoj iteraciji postupka 506 itd., i na taj način dobili cjelokupni popis.



Slika 20: *Rješavanje problema ograničenoga preuzimanja rezultata*

Osim riječi i vrsta riječi, u korpusu možemo pretraživati i riječi koje počinju ili završavaju određenim slovom ili kombinacijom slova (riječi koje počinju slovom 'š' – što, *škola*, *šest*, *šport*, šteta itd.; riječi koje završavaju na *-ost*: *vrijednost*, *mogućnost*, *djelatnost* itd.[89]), što može biti korisno za potrebe poučavanja hrvatskoga jezika.

Pretragu imenskih skupina u korpusu hrvatskoga jezika provela je Borucinsky (2015) na ondašnjemu trećem nacionalnom korpusu u povijesti, i jedinom tada javno dostupnome korpusu - *HNK*, te se rezultati koje je dobila poduda-raju s onima koje smo dobili iz korpusa *hrWaC* i *Riznica.* Rezultati pokazuju

[89] Riječi su poredane prema frekvenciji u korpusu *ENGRI.*

da je neodređena zamjenica (engl. *indefinite pronoun*) najčešća zamjenica u korpusima *hrWaC* i *Riznica*. Međutim, taj podatak valja uzeti sa zadrškom. Naime, poznato je da morfosintaktičko označavanje korpusa nije u potpunosti točno (v. 2.) budući da se obično vrši automatski. Tako su, primjerice, *koji*/*koja*/*koje* označeni samo kao neodređene zamjenice, a ne i kao odnosne zamjenice (pretraživanjem obrasca [tag="Pi.*"] dobit ćemo, između ostalih, i *koji*/*koja*/*koje*, dok se pretraživanjem obrasca [tag="Pr.*"] u *HNK*-u dobije samo zamjenica što, dok ista pretraga u *hrWaC-u*, *Riznici* i *ENGRI-ju* ne daje nikakve rezultate). Nadalje, svako je označavanje rezultat pojedine gramatičke teorije ili teorijskoga pristupa prema kojemu se izvodilo, bez obzira na to što bi ono u biti trebalo biti neutralno ili što neutralnije. Budući da sintaktički pristupi u hrvatskome jezikoslovlju nisu u potpunosti razrađeni i da se mnogi miješaju, posljedica je nepravilno morfosintaktičko označavanje. To je vrlo ozbiljan problem koji u pitanje dovodi rezultate pretraživanja prema morfosintaktičkim kategorijama.

Tablica 8: *Zamjenice u hrvatskome jeziku prema* hrWaC-u *i* Riznici 1.0

Zamjenica	**CQL**	**Primjer**	**RF hrWaC**	**RF Riznica**	**RF HNK (2015)**
neodređena zamjenica (indefinite pronoun)	[tag ="Pi.*"]	neki, ništa, čemu	22744,41	19629,53	17154,4
povratna zamjenica (reflexive pronoun)	[tag ="Px.*"]	svojim, se	19556,16	18545,51	14785,1
osobna zamjenica (personal pronoun)	[tag="Pp.*"]	vas, nam, ga	18413,63	12356,78	5640,1
pokazna zamjenica (demonstrative pronoun)	[tag ="Pd.*"]	ovim, ova, tu	16294,85	13695,88	11380,8
posvojna zamjenica (possessive pronoun)	[tag ="Ps.*"]	naš, njegov, vaš	4774,81	4355,07	3236,1
upitna zamjenica (interrogative pronoun)	[tag ="Pq.*"]	tko, što, kakve	176,76	216,99	0,8
odnosna zamjenica (relative pronoun)	[tag ="Pr.*"]		0	0	0?

Do sada smo prikazali kako pronaći frekvencije za vrste riječi za koje je korpus označen, odnosno za unaprijed definirane riječi ili jezične kategorije. No, crpljenje i pronalaženje cjelokupnoga popisa ili kategorije kao što su neprilagođene engleske riječi i njihovih frekvencija u hrvatskome znatno je složenije. Budući da oznaka isključivo za strane riječi ne postoji u korpusu, tj. da riječi u pravilu nisu označene kao strane, ovo naočigled jednostavno pitanje pokazalo se vrlo složenim. Stoga se problemu pristupilo iz nekoliko perspektiva, a te su metode opisane u nastavku.

6. CRPLJENJE ENGLESKIH RIJEČI IZ KORPUSA HRVATSKOGA JEZIKA

U poglavlju 4.4. prikazali smo načine izrade korpusa za potrebe istraživanja engleskih riječi u hrvatskome. U tome smo poglavlju definirali što smatramo engleskim riječima, a u ovome se poglavlju bavimo crpljenjem engleskih riječi i njihovih frekvencija iz korpusa hrvatskoga jezika. Motivacija je ovoga poglavlja da se usporede rezultati crpljenja riječi iz korpusa s drugim metodama, npr. algoritamskom klasifikacijom koja je primjenjena u projektu *Engleske riječi u hrvatskome* (usp. Bogunović, I. & Kučić, M., u postupku recenzije).

U svrhu pronalaženja engleskih riječi u drugim su se jezicima koristile različite metode. Engleske se riječi sporadično javljaju u hrvatskim tekstovima, stoga je potreban velik niz podataka i preciznija granularnost (Hughes i dr., 2006) prilikom analize i crpljenja takvih riječi (Serigos, 2017: 25). Modeli za (fino-granularno) otkrivanje jezika koriste jedan od sljedećih pristupa, ili pak kombinaciju tih pristupa: tzv. *char-grami*, uparivanje uzoraka (engl. *pattern matching*) i tehnike traženja riječi (engl. *look-up techniques*) (Serigos, 2017: 20). Prvi pristup temelji se na pretpostavci da jezike čine jedinstvena slova/znakovi ili niz slova/znakova, a takvi pristupi zahtijevaju velike korpuse koji će poslužiti kao korpusi za uvježbavanje koji računaju vjerojatnost da će se niz slova pojaviti u danome jeziku (Furiassi i Hofland, 2007 u Facchinetti, 2015; Furiassi, 2008; Serigos, 2017). Drugi model također pretražuje specifične obrasce koji se, za razliku od prvoga modela, unaprijed zadaju. Naposljetku, tehnike pretraživanja koriste se resursima kao što su rječnici, liste riječi i druga referentna djela (Alex, 2005).[90] Borucinsky i Bogunović (2022) daju pregled različitih metoda crpljenja engleskih riječi koje su primjenjivane u drugim jezicima. Tako kao primjer navode istraživanje koje je provela Núñez Nogueroles (2016) u kojem koristi popis engleskih riječi koje pretražuje u nacionalnome korpusu španjolskoga jezika. Drugi je primjer istraživanje iz područja sporta (Balteiro, 2011) gdje se riječi iz engleskoga jezika prikupljene iz rječnika anglizama pretražuju u nacionalnome korpusu u cilju dobivanja podataka o njihovoj učestalosti. Kako autorice (ibid.) navode, druge metode crpljenja engleskih riječi uključuju razvoj novih računalnojezikoslovnih alata i/ili resursa. Na primjer, pod pretpostavkom da broj rezultata dobivenih *Googleovim* pretraživanjem može

[90] Ovo je također opisano u Jelčić Čolakovac i Borucinsky (u tisku).

pokazati pripadnost jeziku razvijen je nenadzirani sustav za prepoznavanje engleskih riječi u njemačkome (koji je primijenjen i na francuski) koristeći postojeće leksičke baze i mrežno dostupne podatke (Alex, 2005). Takav pristup može biti problematičan za jezike koji su nedovoljno zastupljeni na internetu, kao što je slučaj s hrvatskim. Mrežno pretraživanje može se izbjeći pretraživanjem leksikona u kombinaciji s n-gramima (npr. Furiassi i Hofland, 2007 u Facchinetti, 2015), no ta metoda također neće odgovarati jezicima s nedovoljno razvijenim resursima, poput hrvatskoga. Ti se problemi mogu izbjeći pomoću metoda nadziranoga učenja (npr. Castro i dr., 2017; Losnegaard i Lyse, 2012; Serigos 2017). No, takve metode zahtijevaju označene podatke za treniranje klasifikatora, što znači da u slučaju da takvi podaci ne postoje, treba ih izraditi (usp. Bogunović i Kučić, u postupku recenzije). Jedan je od značajnijih problema činjenica da se korpusi dugo izrađuju i da je potrebno uložiti značajna financijska sredstva i vrijeme istraživača za izradu i ažuriranje korpusa. A u trenutku kada se provode istraživanja, korpus je već zastario pa u njemu ne nalazimo neologizme.

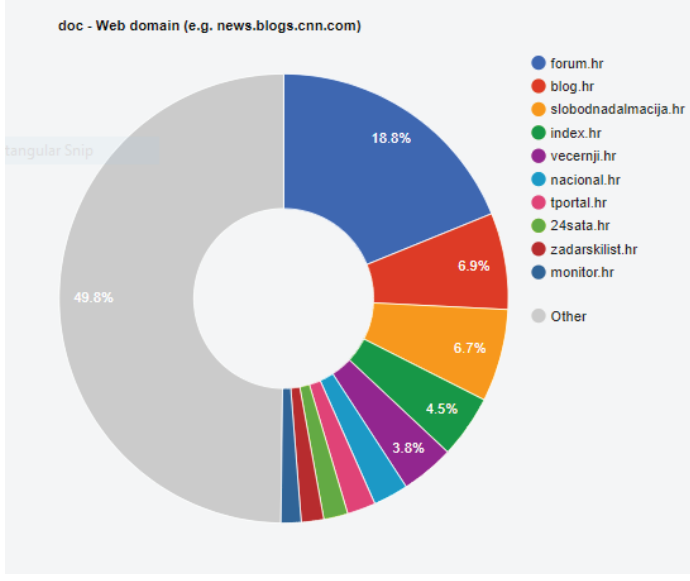
6.1. METODA BR. 1 – OZNAKA *XF*

Jedan je od načina za pronalaženje engleskih riječi u korpusu hrvatskoga jezika pretraga korpusa pomoću CQL-a, odnosno oznake *Xf*[91] ([tag="Xf"], gdje je X − *residual* (preostalo), f − *foreign* (strano)[92]) u računalnojezikoslovnome alatu *SkE* ili *NoSkE* u mrežnome korpusu hrvatskoga jezika (*hrWaC*).

Kako bi se dobio uvid u najveće mrežne stranice u *hrWaC*-u, kao i one s najvećim brojem riječi koje su označene kao strane riječi, uspoređeni su popisi domena prema količini teksta bez kriterija oznake *Xf* i s kriterijem oznake *Xf*. Pritom su izdvojene mrežne stranice s najvećim brojem riječi označenih kao *Xf*. Takva analiza, odnosno analiza vrste teksta pokazala je da su najveće mrežne stranice zastupljene u *hrWaC*-u sljedeće: *forum.hr*, *blog.hr*, *slobodnadalmacija.hr*, *index.hr i vecernji.hr*, a čine oko 40 % svih mrežnih stranica zastupljenih u *hrWaC*-u (slika 21 i 22).

[91] Ovo je također opisano u Borucinsky i Bogunović (2022).

[92] V. *MULTEXT-East Croatian part-of-speech tagset (version 5)*.

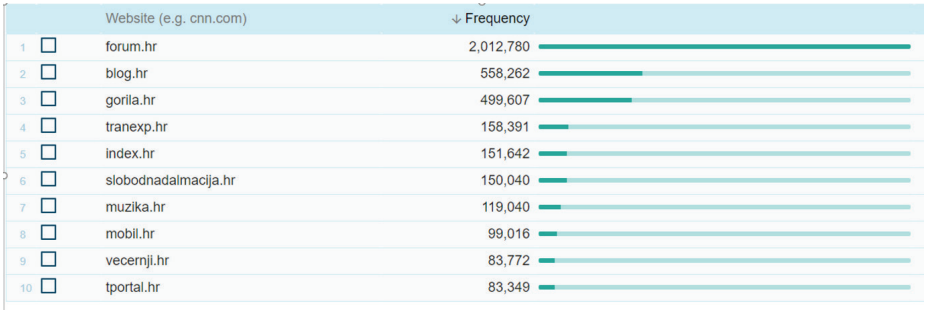


Slika 21: *Najveće mrežne stranice zastupljene u* hrWaC-u

	Attribute value	Token coverage ?		Attribute value	Token coverage ?		Attribute value	Token coverage ?
1	forum.hr	241,694,709	11	politika.com	15,135,831	21	glas-koncila.hr	8,745,629
2	blog.hr	89,155,427	12	ezadar.hr	14,930,127	22	mojblog.hr	8,573,347
3	slobodnadalmacija.hr	85,486,775	13	gorila.hr	12,996,711	23	bug.hr	8,492,757
4	index.hr	58,088,753	14	poslovni.hr	12,900,673	24	business.hr	8,052,473
5	vecernji.hr	49,237,340	15	net.hr	12,184,209	25	hidra.hr	7,881,502
6	nacional.hr	32,353,100	16	dnevnik.hr	11,712,536	26	vidi.hr	7,509,243
7	tportal.hr	26,371,865	17	bloger.hr	11,538,448	27	skole.hr	7,481,005
8	24sata.hr	22,306,953	18	jutarnji.hr	9,772,792	28	advance.hr	6,606,861
9	zadarskilist.hr	20,746,707	19	mobil.hr	9,299,934	29	glasistre.hr	6,526,600
10	monitor.hr	18,495,268	20	matica.hr	9,192,573	30	057info.hr	6,436,381

Slika 22: *Raspodjela pojavnica u korpusu* hrWaC *prema vrsti teksta*

Nadalje, najveći broj riječi označenih kao *Xf* nalazimo na mrežnim stranicama *forum.hr*, *blog.hr*, *gorila.hr*, *tranexp.hr*, *index.hr* itd. kao što je prikazano na slici 23, što bi moglo navesti na zaključak da se na forumima i u medijima češće pojavljuju neprilagođene engleske riječi u odnosu na druge vrste tekstova. No, s obzirom na nepouzdanost ove vrste pretraživanja i s obzirom na to da oznaka *Xf* nije primarno namijenjena crpljenju stranih riječi, ovu tvrdnju valja uzeti sa zadrškom.



	Website (e.g. cnn.com)	↓ Frequency
1	forum.hr	2,012,780
2	blog.hr	558,262
3	gorila.hr	499,607
4	tranexp.hr	158,391
5	index.hr	151,642
6	slobodnadalmacija.hr	150,040
7	muzika.hr	119,040
8	mobil.hr	99,016
9	vecernji.hr	83,772
10	tportal.hr	83,349

Slika 23: *Mrežne stranice s najvećim brojem riječi označenih kao* Xf *u* hrWaC*-u*

U sljedećemu koraku uslijedilo je crpljenje stranih riječi iz korpusa hrvatskoga jezika preko liste riječi (v. 3.) te pregledavanje i ručno pročišćavanje rezultata usporedbom dobivenih konkordancija (v. 3.). Na temelju te usporedbe i dobivenih rezultata uslijedila je ručna klasifikacija riječi prema kriterijima navedenima u 4.4. S obzirom na to da je pretraga osjetljiva na vrstu upita i postavke pretrage (v. 1.8.), provedena je analiza koje će postavke dati optimalne rezultate, odnosno uspoređeni su rezultati pretrage prema lemi, obliku riječi[93] te postavkama velika/mala slova. Kako bi se optimizirala pretraga, podešen je kriterij pretrage na sve riječi koje su u korpusu označene kao *Xf* i sadrže više od tri znaka, tj. [word=".{3,}" & tag="Xf"].

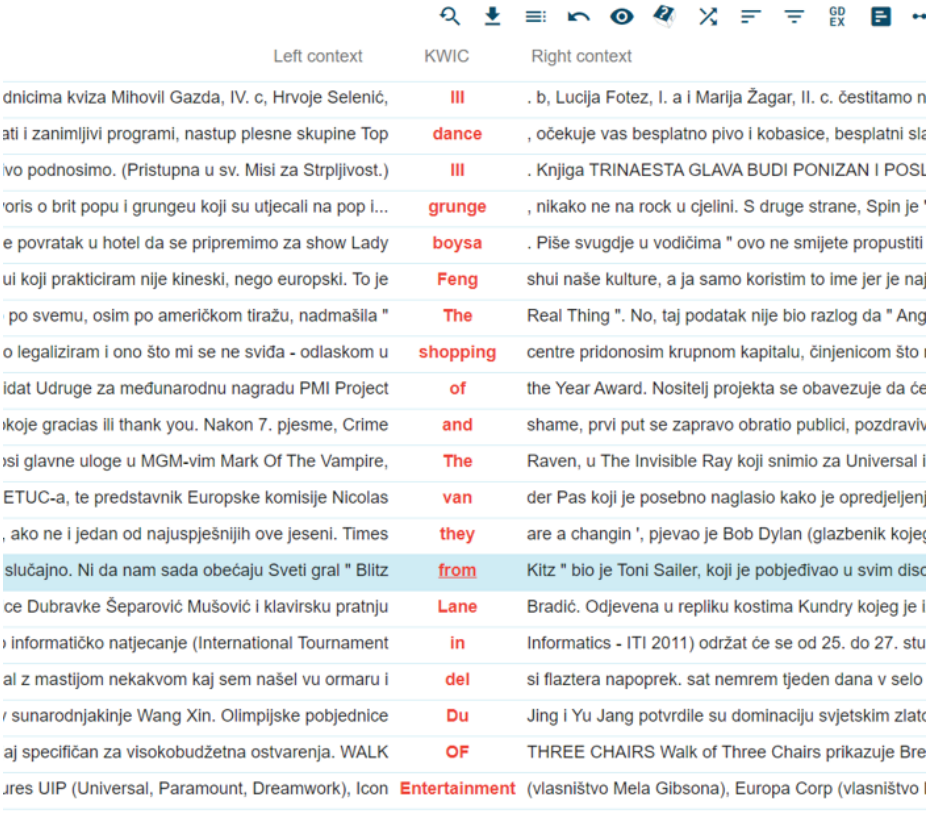
Utvrđeno je da će se riječi pretraživati prema obliku riječi tako da se obuhvate riječi *time* i *times,* i da pretraga neće biti osjetljiva na velika i mala slova, odnosno da će obuhvatiti riječi napisane i velikim i malim slovima. S tim postavkama pomoću oznake *Xf* od ukupnoga je broja pojavnica u korpusu *hrWaC* (1 405 794 913) izdvojeno 6 719 932 stranih riječi, odnosno 4780,17 pojava na milijun riječi. Rezultat od 6,7 milijuna pojavnica sadrži veliku količinu šuma te ga je stoga valjalo pročistiti. Na slici 24 prikazano je prvih 20 konkordancijskih nizova pretrage [tag="Xf"].

[93] Tako se primjerice za postavku pretrage lema dobije 2209 unosa, a za postavku oblik riječi 2976 unosa.

Left context	KWIC	Right context
›odručje za ptice, temeljem europskih i nacionalnih kriterija, te programa	NATURA	2000. Osim ptica, ugroženih na nacionalnoj razini, na ovom području obit
ke Direktive o pticama, koja predstavlja jedno od temeljnih načela izrade	NATURA	2000 mreže u Hrvatskoj. Tako na području Krapinsko zagorske županije o
dokument svog vremena, djelujući svojom poezijom na mnoge pjesnike	XX	. stoljeća. Ostali književnici djeluju pretežito na regionalnoj osnovi, pa se t
pisac Hrvatskog zagorja i oštar kritičar zagrebačkih prilika, čije je djelo "	U	registraturi " najvrednije djelo hrvatskog realizma. iz Hrvatskog zagorja je
uniji, uz Konvenciju za zaštitu financijskih interesa Europskih zajednica (	SL	C 221, 19. 7. 1997., str. 12.). Primjena kaznenog zakonodavstva Članak 2
ji, uz Konvenciju za zaštitu financijskih interesa Europskih zajednica (SL	C	221, 19. 7. 1997., str. 12.). Primjena kaznenog zakonodavstva Članak 2. A
ı iz Gornje Jelenske krećemo prema najvišem vrhu Kalnika Vranilac 643	m	n / v pa putom preko sedam zubi koji vodi grebenom. Atraktivan je, pruža
: Gornje Jelenske krećemo prema najvišem vrhu Kalnika Vranilac 643 m	n	/ v pa putom preko sedam zubi koji vodi grebenom. Atraktivan je, pruža lij
u kako se provlači kroz stijene. Obišli smo i Kalničku gradinu a potom je	ex	predsjednik društva častio sa grahom povodom svog rođendana. SLIKE U
na okolne planine i naš slijedeći cilj, Županj vrh (1138 m). Od Brajkovog	do	Županj vrha stiže se laganom markiranom stazom kroz šumu te se na sar
em MBS: 080230965. Račun društva se vodi kod Erste Steiermärkische	Bank	d. d. u Zagrebu, broj: 2402006 - 1100048308. Temeljni kapital društva izn
ı je okupila četiri medija - televizije NIT i HRT, Radio Istru te Glas Istre, u	gastro	showu pobijedila je ekipu političara načelnike Bala, Žminja, Pićna i Gračiš
e Darwin now - Darwin danas, (organizatori: Prirodoslovni muzej Rijeka,	British	Council) otvorene 20. listopada 2009. u prostoru HKD u Rijeci. Izložbom j
in now - Darwin danas, (organizatori: Prirodoslovni muzej Rijeka, British	Council	) otvorene 20. listopada 2009. u prostoru HKD u Rijeci. Izložbom je bio da
ebu i u Rijeci, 19. studenog 2009. g. Predavanje pod naslovom " Darwin	and	Wallace održao je dr. sc. Paul White (University of Cambridge), nakon čeg
vanje pod naslovom " Darwin and Wallace održao je dr. sc. Paul White (	University	of Cambridge), nakon čega je održan znanstveni kafić uz sudjelovanje pu
naslovom " Darwin and Wallace održao je dr. sc. Paul White (University	of	Cambridge), nakon čega je održan znanstveni kafić uz sudjelovanje publi
aslovom " Darwin and Wallace održao je dr. sc. Paul White (University of	Cambridge	), nakon čega je održan znanstveni kafić uz sudjelovanje publike iz sva tri
će Vam izaći u susret s uslužnošću i izuzetno profesionalnom uslugom.	Wave	Boat Sealver, francuska tvrtka u suradnji sa Yamaha Motor Europa napra
sportskih skutera već 10 godina. Yamaha za 2011. godinu lansira novi "	Special	Edition " model - T-Max Tech Max. Nabrojili smo vam TOP - 5 u Yamahinc

Slika 24: *Konkordancije dobivene pretragom* [tag="Xf"] *u korpusu* hrWaC

Pregled konkordancijskoga niza iz slike 24 pokazuje da dobiveni rezultati ne odgovaraju definicji i kriterijima engleskih riječi postavljenima u 4.4. jer su: 1. iz stranoga jezika koji nije engleski (*NATURA*); 2. vlastita imena ili toponimi (*Wave Boat Sealver, Cambridge*), 3. lažni anglizmi (*gastro show*), 4. pogrešno označene riječi (hrvatski prijedlog *do* označen je kao strana riječ) i sl. Rezultati dobiveni primjenom funkcionalnosti slučajni uzorak (v. 3.) nešto su bolji, npr. riječi poput *grunge*, *dance*, *boysa*, *shopping* itd. spadale bi u kategoriju engleskih riječi, no i ovaj popis sadrži previše šuma, npr. vlastita imena, gramatičke riječi i sl. (slika 25).



Left context	KWIC	Right context
dnicima kviza Mihovil Gazda, IV. c, Hrvoje Selenić,	III	. b, Lucija Fotez, I. a i Marija Žagar, II. c. čestitamo n
ati i zanimljivi programi, nastup plesne skupine Top	dance	, očekuje vas besplatno pivo i kobasice, besplatni sla
ivo podnosimo. (Pristupna u sv. Misi za Strpljivost.)	III	. Knjiga TRINAESTA GLAVA BUDI PONIZAN I POSL
oris o brit popu i grungeu koji su utjecali na pop i...	grunge	, nikako ne na rock u cjelini. S druge strane, Spin je '
e povratak u hotel da se pripremimo za show Lady	boysa	. Piše svugdje u vodičima " ovo ne smijete propustiti
ui koji prakticiram nije kineski, nego europski. To je	Feng	shui naše kulture, a ja samo koristim to ime jer je naj
po svemu, osim po američkom tiražu, nadmašila "	The	Real Thing ". No, taj podatak nije bio razlog da " Ang
o legaliziram i ono što mi se ne sviđa - odlaskom u	shopping	centre pridonosim krupnom kapitalu, činjenicom što
idat Udruge za međunarodnu nagradu PMI Project	of	the Year Award. Nositelj projekta se obavezuje da će
koje gracias ili thank you. Nakon 7. pjesme, Crime	and	shame, prvi put se zapravo obratio publici, pozdraviv
si glavne uloge u MGM-vim Mark Of The Vampire,	The	Raven, u The Invisible Ray koji snimio za Universal i
ETUC-a, te predstavnik Europske komisije Nicolas	van	der Pas koji je posebno naglasio kako je opredjeljenj
, ako ne i jedan od najuspješnijih ove jeseni. Times	they	are a changin ', pjevao je Bob Dylan (glazbenik koje
slučajno. Ni da nam sada obećaju Sveti gral " Blitz	from	Kitz " bio je Toni Sailer, koji je pobjeđivao u svim disc
ce Dubravke Šeparović Mušović i klavirsku pratnju	Lane	Bradić. Odjevena u repliku kostima Kundry kojeg je i
informatičko natjecanje (International Tournament	in	Informatics - ITI 2011) održat će se od 25. do 27. stu
al z mastijom nekakvom kaj sem našel vu ormaru i	del	si flaztera napoprek. sat nemrem tjeden dana v selo
sunarodnjakinje Wang Xin. Olimpijske pobjednice	Du	Jing i Yu Jang potvrdile su dominaciju svjetskim zlat
aj specifičan za visokobudžetna ostvarenja. WALK	OF	THREE CHAIRS Walk of Three Chairs prikazuje Bre
ures UIP (Universal, Paramount, Dreamwork), Icon	Entertainment	(vlasništvo Mela Gibsona), Europa Corp (vlasništvo

Slika 25: *Rezultati pretraga nasumičnih rječničkih primjera*

Funkcionalnost lista riječi (v. 3.) izlistava frekvencije riječi u korpusu kojima je dodijeljena oznaka *Xf*, no kao što je razvidno iz slike 26 dobiveni rezultati nisu odgovarajući jer su neke hrvatske riječi (*taj*, *on*) označene kao strane, a ostale riječi s popisa uglavnom obuhvaćaju gramatičke riječi, koje nisu obuhvaćene ovim istraživanjem.

	Lemma	↓ Frequency	Per million tokens
1	the	665,207	473.19
2	of	356,096	253.31
3	and	259,287	184.44
4	in	238,426	169.60
5	taj	186,379	132.58
6	is	109,874	78.16
7	for	100,471	71.47
8	on	86,882	61.80
9	that	86,613	61.61
10	it	82,980	59.03
11	you	80,667	57.38
12	with	62,346	44.35
13	by	57,608	40.98

Slika 26: *Frekvencije dobivene pretragom* [lemma="Xf"]

Nakon ručne klasifikacije pomoću konkordancije i svođenja na zajedničku lemu, od 2289 riječi preostalo je 1217 riječi, što znači da je 47 % rezultata bio šum. Nadalje, jedan od problema s kojima smo se susreli jest pojavljivanje riječi označenih kao *Xf* u engleskome kontekstu, što je trebalo ručno pregledati i ukloniti takve riječi s popisa, a drugi je problem činjenica da se riječi pojavljuju u identičnome kontekstu, ali u drugome izvoru kao npr. riječ *shared* koja se tri puta pojavljuje u korpusu u identičnim rečenicama, ali je iz drugoga izvora, kao što je prikazano u primjeru (3).

> (3) *Također je odlučeno da se pokrene projekt objedinjavanja informacijskog sustava Hrvatskog zavoda za mirovinsko osiguranje (HZMO) u Shared Service Centru, koji se uspostavlja u Agenciji za podršku informacijskim sustavima i informacijskim tehnologijama* (APIS IT).

Nakon klasifikacije pomoću konkordancija i svođenja na zajedničku lemu, bilo je potrebno završno ručno čišćenje zbog pojave određenoga broja riječi označenih kao *Xf* u engleskome kontekstu, kao i pojave nekih riječi u identičnome kontekstu, ali drugome izvoru, a pročišćeni popis prikazan je u tablici 9.

Tablica 9: *Prikaz riječi označenih kao* Xf u hrWaC-u *nakon ručnoga pročišćavanja*

Red. br.	hrWaC_ riječ (mala slova)	AF	RF
1.	top	38439	2.734.325
2.	world	28334	2.015.514
3.	love	26789	1.905.612
4.	time	24122	1.715.898
5.	more	23996	1.706.935
6.	sex	21261	1.512.383
7.	big	20436	1.453.697
8.	live	20415	1.452.203
9.	times	19535	1.094.043
10.	man	18802	1.337.464
11.	day	18741	1.333.125
12.	open	18630	1.325.229
13.	life	18458	1.312.994
14.	grand	17678	1.257.509
15.	online	16852	1.178.266
16.	like	16453	1.170.370
17.	best	16297	1.159.273
18.	full	16122	1.146.824
19.	do	14866	1.057.480
20.	ad	14649	1.042.044
21.	go	14118	1.004.272
22.	music	14052	955.189
23.	only	14013	996.803
24.	make	13771	979.588
25.	art	13747	977.881
26.	house	13412	902.906
27.	home	13383	951.988
28.	high	12860	914.785
29.	sorry	12533	891.524
30.	new	12203	868.050
31.	hard	12174	865.987
32.	gay	11775	837.604
33.	play	11625	826.934
34.	bad	11608	804.528

35.	daily	11604	825.440
36.	international	11435	813.419
37.	blue	11187	795.778
38.	jam	11156	0.88064
39.	night	11125	791.367
40.	ex	10964	779.915
41.	wall	10920	776.785
42.	flash	10886	774.366
43.	fashion	10688	760.282
44.	sun	10247	728.911
45.	national	10241	728.485
45.	info	10178	724.003

Zaključno, oznaka *Xf* nije pouzdan način crpljenja engleskih riječi u hrvatskome jer nije namijenjena ovoj svrsi. Da bi se pouzdano mogle crpiti engleske riječi iz korpusa hrvatskoga jezika valja razviti korpusni alat ili statističku metodu koja bi omogućila crpljenje takvih riječi, a jedan primjer prikazan je u 6.2.

U nastavku opisujemo nekoliko problema s kojima smo se susreli prilikom izrade *Baze engleskih riječi i hrvatskih istovrijednica* (Bogunović, Jelčić Čolakovac & Borucinsky, 2022), a koji ilustriraju na što valja obratiti pažnju prilikom slanja upita i pretrage, odnosno dohvaćanja i interpretacije podataka iz korpusa. Navedena baza nadogradnja je *Baze engleskih riječi u hrvatskome*, koja je sastavljena pomoću algoritma koji je razvio Kučić, a metoda je opisana u Kučić i Bogunović (2022). U *Bazi engleskih riječi i hrvatskih istovrijednica* izvršena je nadogradnja na način da su dodane neke nove riječi, riječi su podijeljene na jednorječne i višerječne izraze, dodane su frekvencije iz dvaju korpusa, kao i hrvatske istovrijednice. Iako je priprema teksta i crpljenje engleskih riječi za potrebe projekta učinjena algoritamskom klasifikacijom, u konačnici su jezikoslovci ipak morali provjeriti rezultate na način da su pregledavali frekvencije i konkordancije u korpusu. Bez pristupa korpusu to bi bilo jako teško.

Ako u korpusu primjerice tražimo englesku riječ *one* ‘jedan’, a ne želimo dobiti rezultate za hrvatsku zamjenicu ženskoga roda množine *one,* pretragu možemo definirati na sljedeći način: [word="one" & tag="Xf"]. To vrijedi i za primjere poput engleske imenice *van* ‘kombi’ i hrvatskog priloga *van*, engleske imenice *rose* ‘ruža’ i hrvatske imenice *rosa*, engleske imenice *medicine* ‘1. medicina,

2. lijek' te hrvatske imenice *medicina*. Isto tako, tražimo li u korpusu hrvatskoga jezika engleske riječi popu *cast* i *split,* najbolje je definirati pretrage kao [lemma="cast" & tag="Xf"] odnosno [lemma="split" & tag="Xf"]. Međutim, ovakve pretrage izlistat će samo one riječi koje su označene kao strane, tj. one koje je označivač prepoznao. Detaljniji pregled korpusa pokazuje da ima i puno više primjera takvih riječi koje su pogrešno označene kao hrvatske, pa ih se takvom pretragom neće obuhvatiti. Nadalje, kod riječi *cast*, primjerice, nalazimo još jedan problem, a taj su dijakritički znakovi, pa ćemo tu riječ naći u kontekstima u kojima stoji umjesto riječi čast. Primjerice, za englesku riječ *government SkE* je dao 492 rezultata, a iste je trebalo filtrirati na način da se pomoću filtera prikažu samo rezultati u hrvatskome kontekstu te ručno provjere dobiveni rezultati. Osim ovdje navedenih nekoliko primjera i problema kao što su sinkretizam oblika, dijakritički znakovi (*cast* i *čast*), vlastita imena (npr. engleski glagol *split* 'razdijeliti' u odnosu na grad Split), u sastavljanju popisa engleskih riječi u hrvatskome istraživači su se susreli s brojnim drugim poteškoćama kao što je različit stupanj leksikalizacije (npr. *gem* 'dragi kamen' u odnosu na gem 'dio seta u tenisu, gejm'), treba li u taj popis uvrstiti i metonime, kako riješiti višerječne nazive poput *all-inclusive ponuda* te lažne parove poput *glazbeni spot*, *far*, *fan* i dr.

Kako bi se sastavila pouzdana frekvencijska lista bilo je nužno razriješiti sve probleme i ručno provjeriti svaku pojavnicu. Nema univerzalnoga i automatskoga rješenja za takve poteškoće, već je potrebno kombinirati različite upite da bi se u konačnici dobio valjan rezultat.

Nadalje, različite frekvencije dobit ćemo iz konkordancija i lista riječi jer se kod prve pretrage broje frekvencije samo za konkordancije, dok se kod potonje pretrage broje riječi iz cijeloga korpusa. Naravno da će dobiveni rezultati ovisiti o tome kako poimamo riječ, lemu itd. (v. 3.). Nadalje, kao korisnici korpusa ovisni smo o onima koji razvijaju sustave i alate, jer i male preinake mogu utjecati na dobivene rezultate i valjanost podataka.

U konačnici, rezultati dobiveni uporabom postojećih računalnojezikoslovnih alata za dano istraživanje i dobivanje odgovora na pitanje koje su najčešće neprilagođene engleske riječi u hrvatskome nisu zadovoljavajući jer nismo pronašli riječi koje smo očekivali, kao npr. *freelancer*, *influencer*, *celebrity*, tj. 'novije' riječi. Jedan od mogućih razloga jest činjenica da je *hrWaC* prikupljan 2011. i 2013., a druga da oznaka *Xf* ne daje pouzdane rezultate. Pretražimo li,

primjerice, riječ *freelancer* u *hrWaC*-u, vidjet ćemo da je označena kao hrvatska imenica.

Nadalje, ovim načinom pretrage nećemo dobiti riječi poput *host(ati)*, *download(ati)*, *link(ati)*, *freelancer*, odnosno riječi koje se mogu sprezati ili sklanjati te su automatski označene kao hrvatske riječi. U tu kategoriju može se svrstati i glagol *googlati*, koji je problematičan ne samo zbog hrvatskoga infinitivnog nastavka, već i zbog činjenice da je izveden iz naziva *Google* koji se može svrstati pod vlastita imena čime ne bi bio uvršten u popis, a zapravo je značajan. Isto tako, podaci dobiveni analizom ovoga korpusa ne zrcale trenutno jezično stanje, jer korpus nije redovito ažuriran. Stoga se javlja potreba za stvaranjem korpusa koji će ponuditi novije podatke, a koji bi se mogao koristiti u kombinaciji s podacima dobivenima iz *hrWaC*-a. Nadalje, oznaka *Xf* nije primarno zamišljena za crpljenje stranih riječi, već je to kategorija u koju se svrstavaju riječi koje označivač nije prepoznao, a to mogu biti kratice, akronimi, tipfeleri, strane riječi itd. Stoga pretraga preko ove oznake zahtijeva ručno pročišćavanje.

Liste ili popisi riječi i frekvencije najčešće su polazište za daljnja istraživanja (no ne moraju uvijek dati optimalne rezultate), a ovo istraživanje pokazalo je da rezultati znatno variraju. Frekvencije ne ovise samo o tekstovima koji čine korpus, već i o sustavima za obradu korpusa, stoga je potrebno poznavati mogućnosti i ograničenja alata, otkriti razloge za nezadovoljavajuće rezultate i pokušati ih riješiti. Rezultati ovoga istraživanja pokazali su da se postojećim alatima i resursima za hrvatski jezik mogu pronaći neke engleske riječi u hrvatskome, no isto tako da se popis dobiven ovom metodom ne može smatrati cjelovitim, pouzdanim i reprezentativnim. Stoga se postavlja pitanje vrijedi li se baviti korpusima s obzirom na navedene poteškoće. Odgovor na to pitanje je potvrdan jer ovakva analiza ne zahtijeva veliku financijsku potporu (ne uzimajući u obzir postupak sastavljanja korpusa), a puno je brža od ručnoga crpljenja riječi. Nije optimalna, no može se kombinirati s drugim metodama (v. 6.2.). Ipak, za stvaranje baze engleskih riječi čini se potrebnim dopuniti postojeće resurse (korpuse) novijom građom te razviti nove alate koji će preciznije i učinkovitije klasificirati engleske riječi u hrvatskome jeziku. Pored značajnoga doprinosa u istraživanju pojave engleskih riječi u hrvatskome, ovaj istraživački smjer uvelike bi pridonio i razvoju računalnojezikoslovnih alata i resursa u hrvatskome jeziku.

Zaključno, korpusni su alati vrlo moćni, no znatno su bolji za istraživanje tipičnih i karakterističnih jezičnih pojava, a engleske riječi u hrvatskome, premda ih ima popriličan broj, nisu tipične. Stoga ne čudi da alat nije dao zadovoljavajuće rezultate za engleske riječi u hrvatskome. Ako su riječi pak prilagođene kao što je slučaj s rječju *freelancer*, tada ih alat ne smatra iznimkama pa se ovom pretragom neće pronaći takve riječi. Rješenje ovoga problema bilo bi da se u korpusu označe engleske riječi prema definiranim kriterijima, a jedan primjer kako bi se to moglo učiniti prikazan je u sljedećem potpoglavlju.

6.2. METODA BR. 2. – N-GRAMI

Ova metoda temelji se na pronalaženju engleskih riječi na temelju tipičnih kombinacija slova i/ili znakova za dani jezik. Vrlo pojednostavljeno to znači da u hrvatskome, primjerice, nećemo naći kombinacije slova *wh-*, *-ing*, *you-*, *-ess* i sl., dok u engleskome nećemo pronaći slova koja sadrže dijakritičke znakove i sl. Početna ideja bila je pronaći sve jedinstvene kombinacije za engleski i hrvatski jezik te ih postaviti kao osnove za pretragu, odnosno usporediti. U ovome se dijelu prvenstveno fokusiramo na tzv. *char-grame* (engl. *character grams*[94]) ili nizove slova, a riječi kao N-grami opisane su u poglavlju 3. Budući da *SkE* ne podržava pretragu prema tzv. *char-grams* već samo N-grams pri čemu je jedan N-gram riječ, a ne slovo, takva pretraga nije bila moguća u danome alatu. Iz tog razloga korišten jealat *Phrases in English* (Fletcher, 2010), na temelju kojega su dobivene najčešće kombinacije od tri slova za engleski jezik na početnom, središnjem i/ili finalnom položaju s minimalnom frekvencijom od 100 pojavnica i 10 različnica. Kako engleski jezik ima najrazvijenije jezične tehnologije, ovaj podatak dobiven je u svega nekoliko klikova, a 10 najčešćih *char-grama* koji sadrže tri slova prikazano je u tablici 10.

[94] Baza *char-grama* dobivena je iz svih oblika riječi koji se pojavljuju 10 ili više puta u korpusu *BNC* (Fletcher, 2010).

Tablica 10: *Char-grami za engleski jezik*

Red. br.	Char-gram	AF	RF[95]	Položaj (početni, središnji, krajnji)	CQL
1.	ing	2991683	9416	krajnji	[lemma=".*ing"]
2.	ion	1961151	3639	krajnji	[lemma =".*ion"]
3.	ter	1146045	3154	središnji, krajnji	[lemma =".*ter.*|.*ter"]
4.	ati	1113306	3093	središnji	[lemma =".*ati.*"]
5.	ate	880095	2975	središnji	[lemma =".*ate.*"]
6.	tio	1552387	2968	središnji, krajnji	[lemma =".*tio.*|.*tio"]
7.	ent	1754690	2794	središnji, krajnji	[lemma =".*ent.*|.*ent"]
8.	ted	705334	2785	krajnji	[lemma =".*ted"]
9.	ers	771603	2580	središnji, krajnji	[lemma =".*ers|.*ers.*"]
10.	tin	536447	2175	početni, središnji, krajnji	[lemma="tin.*|.*tin.*|.*tin"]

Najprije se za svaki *char-gram* u engleskome kontekstu provjerilo na kojem položaju se u pravilu pojavljuje (na početnom, središnjem ili krajnjem), te je na temelju toga postavljena pretraga. Upit CQL-a naveden je u zadnjem stupcu tablice.

Ova pretraga pokazala je sljedeće:

1. Trigrami koji se frekventno pojavljuju u engleskome mogu se pojavljivati i u hrvatskome (npr. *ted*, *ion*, *ati*, *ate*). Bolje rezultate dat će primjerice trigram *-ing* (npr. *trening*, *shopping*, *holding*, *marketing* itd., ali *mesing*) nego *-tin* (npr. *neistina*, životinja, *pukotina* itd.).
2. Metoda je iznimno dugotrajna jer zahtijeva posebnu pretragu za svaku kombinaciju slova, što je popriličan broj. Nadalje, potrebno je filtriranje engleskoga konteksta za svaku riječ, te naposljetku ručno filtriranje.
3. Ovom metodom obuhvaćene su engleske riječi koje primaju hrvatske sufikse (npr. *downloadanje*), što je bilo problematično s prvom metodom odnosno oznakom *Xf*. Ova metoda polučila je preciznije rezultate od metode u kojoj je rabljena oznaka *Xf*, no kao jedina metoda nije dovoljna, već ju je potrebno kombinirati s ostalim metodama.

[95] Brojevi u stupcu *Relativna frekvencija* zorno ilustriraju Zipfov zakon (v. 3.), odnosno kako broj pojavljivanja (u ovom slučaju niza slova) naglo opada.

4. Ova metoda dobra je za pronalaženje inačica i različitih oblika engleskih riječi (*shopping*, *sopping* itd.), te za pronalaženje neologizama. Mogli bismo za hrvatski uzeti popis sufikasa koje je iz korpusa izvukla Filko (2020) ili popis tvorbenih sufikasa imenica, pridjeva i glagola (Pandžić, 2015) pa ih usporediti s ovima iz engleskoga. Primjerice, Furiassi (2007, 2008) je usporedio kombinacije slova u engleskome i talijanskome, te tražio samo ona u kojima se riječi ne poklapaju i to je izvrsna metoda za detekciju neologizama, što nije pak bio primarni cilj ovoga istraživanja.

Nakon što smo prikazali problematiku crpljenja engleskih riječi i njihovih frekvencija iz korpusa hrvatskoga jezika, u nastavku se pažnja posvećuje odnosima među riječima, odnosno podacima o sintagmatskome odnosu koje možemo dobiti iz korpusa.

7. RELACIJA

Jedan od načina na koji je korpusna lingvistika promijenila naše poimanje jezika jest svijest o povezanosti gramatike i leksika, postulatu koji je 1961. iznio Halliday (1961, 1996).[96] Gramatički obrasci, iako apstraktni sami po sebi, realiziraju se kroz leksičke i gramatičke odabire. Veliki korpusi su uvelike doprinijeli tom razvoju jer pružaju empirijske dokaze o povezanosti leksika i gramatike, posebice u anglosaksonskoj tradiciji (usp. primjerice *pattern grammar*, Hunston i Francis, 2002; *collostructions*, Stefanowitsch i Gries, 2003; *word sketches*, Kilgarriff i Tugwell, 2001; *concgrams,* Cheng i dr. 2006; Cheng i dr., 2009 i dr.). U žarištu su ovoga poglavlja leksičko-gramatički obrasci dobiveni trećom vrstom podataka iz korpusa − relacijom ili odnosom prema drugim jezičnim jedinicama. U ovome se poglavlju traže odgovori na pitanje koji je odnos između X i Y. Pri tome se ne istražuju paradigmatski, već sintagmatski odnosi, tj. u kojem se ko(n)tekstu neka jezična jedinica realizirala.

7.1. KORPUSNO ISTRAŽIVANJE LEKSIKA

U ovome će poglavlju naglasak biti na funkcionalnostima *konkordancija* (engl. *Concordance*), *kolokacija* (engl. *Collocation*), *slučajni uzorak* (engl. *Get random sample*) i *dobri rječnički primjeri* (engl. *Good dictionary examples*) (Kilgarriff i dr., 2008) koji su prije svega namijenjeni leksikografima[97], no mogu se koristiti i u nastavi stranoga jezika (struke), kao što su pokazale Borucinsky i Tominac Coslovich (2021). Nadalje se prikazuju funkcionalnosti alata kao što su *tezaurus* (engl. *Thesaurus*), *skica riječi* (engl. *Word Sketch*), *razlike u skicama riječi* (engl. *Word Difference*), te *crpljenje nazivlja* (engl. *Terminology Extraction*).

7.1.1. ISTRAŽIVANJE NA RAZINI RIJEČI

Prva je jezična jedinica kojom započinjemo istraživanje relacija na temelju korpusa riječ. Za ilustraciju je odabrana riječ *oblak.* Jednostavna pretraga

[96] Halliday (1996) u svome radu koristi naziv leksikogramatika, a definira ga kao spoj gramatike i vokabulara, koji su međusobno neodvojivi i čine jedinstvenu cjelinu. Leksikogramatika je konstrukt za izražavanje. Halliday (1996) često za isti taj pojam koristi i skraćeni naziv gramatika. Takva gramatika (leksikogramatika) slojevita je i obuhvaća semantiku, leksikogramatiku i fonologiju (v. Borucinsky, 2015: 10).

[97] O utjecaju na leksikografiju i primjeni korpusa u leksikografiji v. primjerice Hanks (2012); Fellbaum (2014).

preko opcije *konkordancija* pokazuje da se imenica *oblak* u korpusu *Riznica* pojavljuje 3295 puta, što je njezina apsolutna frekvencija, dok je njezina relativna frekvencija 32,37, odnosno riječ *oblak* se 32,37 puta pojavljuje na milijun riječi u korpusu. Iz konkordancija riječi *oblak* možemo stvoriti rječničku natuknicu, a tu će od koristi biti funkcionalnosti *slučajni uzorak* i *dobri rječnički primjeri*, pri čemu će prva izlistati 200 ili 500[98] nasumičnih primjera. Iz dobivenih primjera i rečenica potom ćemo iščitati najtipičnija značenja riječi. Proučimo li konkordancije imenice *oblak*, naći ćemo dva značenja koja odgovaraju onima predloženima na *Hrvatskome jezičnom portalu* (*HJP*, 2021).

1. *meteor. hidrometeor koji se sastoji od velike nakupine sićušnih kapljica vode i/ili kristalića leda ili obojeg koji slobodno lebdi u zraku, ne dotiče tlo*
2. *pren. a. svaka masa čestica koja se kreće zrakom ili lebdi u zraku [oblak prašine; oblak snijega] b. ono što sliči na oblak (npr. mnoštvo mušica)*

Međutim, nećemo naći značenje imenice *oblak* koje se koristi u domeni računarstva. Druga funkcionalnost koju možemo ispitati jest *dobri rječnički primjeri* (Kilgarriff i dr. 2008), koja je osmišljena tako da pronalazi najtipičnije primjere tražene riječi, a funkcionira na način da rečenice spaja na temelju heurističkoga modela čitljivosti i informativnosti, a prije svega mjeri duljinu rečenice, odbacuje one rečenice koje sadrže nefrekventne riječi i one koje se sastoje od jednoga ili dva znaka koje nije slovo, te anaforične izraze.

7.1.2. ISTRAŽIVANJE FRAZEMA I KOLOKACIJA

Osim za definiranje riječi i pronalaženje adekvatnih primjera riječi u rečenici (kontekstu), iz korpusa možemo dobiti i podatke o kolokacijama i frazemima s imenicom *oblak* poput *crni oblaci*, *gust oblak*, *oblak dima*, *oblak vodene pare*, *prijeteći*, *bijeli oblak*, *zidati kule u oblacima*, *dizati koga/što u oblake* itd. Filipović Petrović i Parizoska (2019) pokazale su kako koristiti korpuse u obradi frazema. Autorice su opisale postupak dobivanja frazema i traženje varijantnih oblika preko kolokacije na način da se traže kolokati neke riječi u rasponu od 5 mjesta lijevo i desno, preko skica riječi, odnosno sažetih prikaza

[98] Broj nasumičnih primjera ovisit će o odabiru korisnika.

kolokacijskog potencijala neke riječi, te filtera kojima se definira supojavljivanje specifičnih riječi u razmaku od 5 mjesta. Također su pokazale kako izraditi frazeme u alatu *Lexonomy* (Měchura, 2017) na temelju čega je izrađen *Hrvatski frazeološki rječnik* (*HFR*).[99]

Usporedbu tradicionalne frazeologije i korpusne frazeologije[100] hrvatskoga jezika provele su Filipović Petrović i Parizoska (2019). Filipović Petrović (2020) pokazala je da je varijante i modifikacije frazema moguće istraživati jedino na velikom broju primjera stvarne upotrebe kakav sadržavaju računalni korpusi. Filipović Petrović (2020: 161) nadalje zaključuje da „računalni korpusi daju nebrojene mogućnosti istraživanja tog jezičnog fenomena, a frazeologija i korpusna lingvistika tek trebaju dosegnuti svoj puni potencijal istraživanja". Filipović Petrović (2020) dokazuje da računalni korpusi: 1) donose podatke o postojanju frazema, odnosno učestalosti njegove upotrebe; 2) putem primjera stvarne upotrebe olakšavaju razumijevanje značenja frazema i daju uvid u mogući specifični kontekst u kojem se frazem češće ili jedino javlja; 3) pokazuju sistematičnost u upotrebi varijanata određenog frazema; 4) pokazuju kreativnu upotrebu frazema u komunikacijske svrhe u obliku namjernih modifikacija, te smatra da se korpusi „opravdano mogu smatrati neizostavnim jezičnim alatima u frazeološkim istraživanjima i frazeografskoj praksi" (Filipović Petrović 2020: 161).

Funkcionalnosti dostupne u alatu *SkE* omogućuju pretragu frazema kroz frekvencije, konkordancije i kolokacije. Primjerice, usporedbom frekvencija frazema *biti u oblacima* i *živjeti u oblacima*[101] utvrđujemo da se frazem živjeti u oblacima dvostruko češće pojavljuje u korpusu *Riznica* (14 pojava u odnosu na 7 pojava frazema *biti u oblacima*). Nadalje, usporedimo li kolokacije

[99] Vidi također *Bazu frazema hrvatskoga jezika* koja sadržava frazeme suvremenoga hrvatskog jezika koji su svrstani pod svakom frazemskom sastavnicom, a temelji se na podacima dobivenima iz *Kolokacijske baze hrvatskoga jezika*.

[100] O frazeologiji engleskoga jezika i dostignućima korpusne lingvistike u frazeologiji v. Gray i Biber (2015).

[101] Ovi rezultati dobiveni su sljedećim pretragama: [!tag="V.*"] [tag="Va.*"] [lemma="u"] [lemma="oblak"], što se može iščitati kao: „Pronađi sve oblike leme *oblak* kojoj prethode lema *u* i pomoćni glagol ispred kojega nije drugi glagol.", kako bi se izbjegli rezultati poput *letjeli smo u oblacima*. Sljedeća pretraga [lemma="živjeti"] [lemma="u"] [lemma="oblak"] znači: „Pronađi sve frazeme s varijantim oblicima glagola živjeti (npr. živjeli, *živi*, živi itd.), iza kojega slijede lema *u* te paradigma imenica *oblak* (npr. *oblacima*)."

imenice *oblak,* vidimo da su njezini najčešći i najznačajniji kolokati imenica *prašina (oblak prašine*) te prijedlog *nad (oblak nad).*[102]

Tablica 11: *Kolokacije i statistički podaci supojavljivanja imenice* oblak *u korpusu* Riznica

Kolokat	Frekv.	Kol. frekv.	T-mjera	MI	logDice
prašine	120	975	10,9516	11,8925	9,84687
dima	96	850	9,79515	11,7685	9,56781
nebu	71	1368	8,42089	10,6468	8,9627
crni	52	1610	7,20387	9,96245	8,4404
nebo	57	2564	7,53884	9,42357	8,31645
tamni	29	251	5,38366	11,8013	8,066
bijeli	35	2005	5,90511	9,07475	7,75751
teški	23	1219	4,7876	9,18694	7,38337
magle	19	574	4,35464	9,99788	7,33018
vjetar	29	2701	5,36893	8,37356	7,3082
sunce	31	3137	5,54952	8,25388	7,30314
sivi	17	270	4,12099	10,9255	7,28778
nad	122	22611	10,9791	7,38086	7,26974
prolom	15	24	3,87278	14,2368	7,21036
neba	24	2017	4,88565	8,52182	7,20992
niskim	17	637	4,1181	9,68717	7,14642
tmasti	14	25	3,74144	14,0784	7,11039
nebom	16	566	3,99542	9,7702	7,08524
bijelim	18	1114	4,23414	8,96325	7,06369

Osim za pronalaženje definicija, primjera riječi u rečenici i frazema, jedna od bitnijih primjena korpusa u jezikoslovnim istraživanjima zasigurno je pronalaženje kolokacija. Pitanje kolokacija usko je vezano uz frekvencije. Deset najčešćih riječi u većini su korpusa iste, a upravo se taj podatak rabi da bi se podaci dobiveni frekvencijama bolje iskoristili. U korpusnoj lingvistici kolokacije se definiraju kao sustavno supojavljivanje riječi u uporabi, odnosno riječi koje se češće pojavljuju jedna uz drugu, i kao posljedica toga međusobno utječu na značenje.

[102] Raspon za traženje kolokata je 6 (3L/3R).

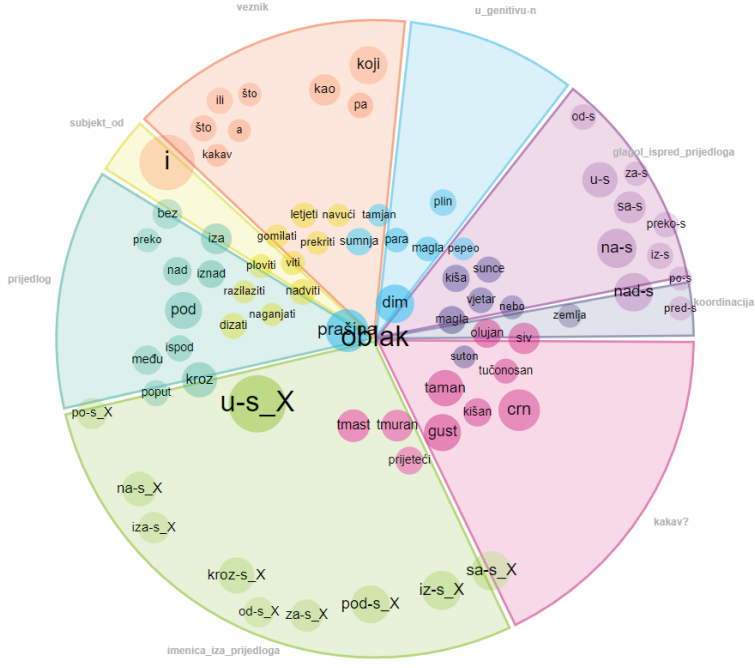
Žustra rasprava vodila se oko kolokacija do kojih bismo došli intuicijom i onih dobivenih iz korpusnih podataka. Uzmemo li za primjer riječ *dama*, vjerojatno će nam na pamet pasti kolokacija *prva dama*, ali se možda nećemo sjetiti kolokacije *damin gambit.* Vrijednost korpusa leži upravo u činjenici da nam omogućuje da na temelju velike količine teksta pronađemo i manje česte ili prototipne kolokacije. U pravilu se kod pretrage kolokacija promatraju kolokati u rasponu od 5 riječi lijevo i desno (5L/5R), no katkada će taj raspon biti uži, primjerice, ako želimo pretraživati kolokacije unutar rečenice.[103] Nije neophodno uvijek promatrati raspon lijevo i desno, za pojedina istraživanja (kada, primjerice, tražimo subjekt glagola) promatrat ćemo samo raspon lijevo od glagola (Lindquist, 2009). Što veći raspon postavimo, to je manja značajnost dobivenoga rezultata (Sinclair, 2004: xxvii). Osim raspona, bitna je i frekvencija (su)pojavljivanja dane riječi koja se obično postavlja na 10 pojava na milijun riječi, iako u literaturi postoje brojne rasprave o tome koliko supojavljivanja je dovoljno da kombinaciju riječi smatramo kolokacijom (usp. Biber i Reppen, 2015). Iz tog razloga valja proučiti statističke podatke koje možemo dobiti iz korpusa te ih prilagoditi kako bismo dobili valjan rezultat, kao što je, primjerice, učinjeno za kolokacije riječi *oblak* prikazane u tablici 13 gdje se raspon od 3 riječi lijevo i desno (3L/3R) pokazao pouzdanijim. U tablici 13 prikazani su statistički podaci na temelju kojih se pronalaze kolokati riječi. Prvi stupac nazvan *frekvencija* prikazuje broj pojavljivanja kolokata (*oblak*) u zadanome rasponu. Drugi stupac *kolokacijska frekvencija* pokazuje koliko se puta kolokat pojavljuje u cjelokupnome korpusu (ili potkorpusu), dok statističke mjere *T-mjera, MI* i *logDice* pokazuju koliko su usko povezani kolokati s kolokatorom (o mjerama supojavljivanja v. 3., usp. također Březina (2018: 72)). Osim navedenih postavki, moguće je u *SkE*-u dobiti i sljedeće statističke podatke o supojavljivanju riječi: *MI3*, *log izglednost*, *min sensitivity*, *MI.log* (Cumming, 2014; Jenset, 2008; Kilgarriff i dr., 2004, Kilgarriff i dr., 2015b; Lijffijt i dr., 2016; Pedersen, 1998; Wallis, 2020).

Frekvencija i log izglednost prigodne su za gramatičke riječi jer ističu frekventne i neekskluzivne kolokate, dok MI i *Log dice* ističu ekskluzivne kolokate.

[103] Osim preko funkcionalnosti *konkordancija*, a potom *kolokacija*, moguće je tražene kolokacije upisati u tražilicu u obliku fraze, iako je autoričino iskustvo pokazalo da takve pretrage funkcioniraju dobro za engleski jezik, a slabije za hrvatski zbog morfološke složenosti hrvatskoga jezika. U tražilicu možemo primjerice upisati englesku frazu *mind business* što će izlistati rezultate poput *mind your own business*, no upisivanje hrvatske fraze *glava oblak* neće polučiti rezultate. Kao i u ostalim pretragama, upiti pomoću CQL-a polučit će najbolje rezultate.

Izbor mjere ovisi o tome koje kolokacije pretražujemo (usp. Gablasova i dr., 2017; McEnery i dr., 2019).

Nadalje, preko skice riječi možemo utvrditi gramatičke odnose među riječima, te ih na jednostavan i praktičan način vizualizirati (slika 25). Udaljenost riječi od središta prikazuje njezinu prototipnost, što je izraženo T-mjerom (npr. kolokacija *oblak prašine* je prototipnija od kolokacije *oblak dima*), dok veličina kruga prikazuje frekvenciju, pa je tako *tamni oblak* učestaliji u korpusu *Riznica* u odnosu na kolokaciju *sivi oblak*. Boje na grafičkome prikazu prikazuju gramatičku kategoriju kojoj riječ pripada.



Slika 27: *Vizualizacija skice riječi* oblak

Osim kolokacija, iz korpusa možemo dobiti podatke i o tzv. koligacijama, odnosno formalnim odnosima među riječima koji pokazuju gramatičku sustavnost u jeziku, npr. *nad*, *iznad*, *kroz*, *iza*, *ispod* + *oblak*, a ti podaci važni su, između ostaloga, u podučavanju jezika. Osim naziva koligacija rabi se i naziv gramatička kolokacija koja pokazuje kako se pojedine imenice češće supojavljuju s određenim prijedlozima, dok kod leksičkih kolokacija kombinacije riječi

imaju određeni značenjski kontekst (npr. *putnički brod*, *trgovački brod*, *zračna luka* itd.) (Posavec, 2017).

Kolokacijski obrasci mogu pokazati i konotacije koje neke riječi imaju, primjerice, riječ *nezaposlen* imat će negativnu konotaciju ili negativnu semantičku prozodiju, što možemo iščitati iz konteksta (npr. *siromašan i nezaposlen*). S druge će pak strane neke riječi imati pozitivnu semantičku prozodiju (npr. *Za svoj **neumorni**, istraživački i nadasve profesionalni novinarski rad B. R. dobitnik je nekoliko društvenih priznanja, od kojih posebno vrijedi istaknuti Medalju grada Varaždina*).

Kolokacije nisu glavna tema ove knjige, iako bi im zasigurno valjalo posvetiti više prostora. U ovoj knjizi će se, na temelju dosadašnjih provedenih istraživanja autorice (usp. Šnjarić i Borucinsky, 2020), prikazati mogućnosti i funkcionalnosti računalnojezikoslovnih alata za istraživanje ove jezične pojave. Za potrebe istraživanja transdisciplinarnih glagolsko-imeničkih kolokacija koje su dijelom ograničene skupine općeznanstvenoga leksika ili onoga što Fandrych (2006) naziva *Textformulierungsroutinen*, odnosno „rutinskim obrascima oblikovanja znanstvenoga teksta“, autorica je sastavila korpus *Općeznanstvenih kolokacija hrvatskoga jezika* (*OZJ*)[104] koji čine znanstveni članci humanističkih znanosti objavljeni od 2015. do 2017. godine, dostupni na mrežnome portalu *Hrčak,* te disertacije preuzete iz digitalnoga repozitorija *Nacionalne i sveučilišne knjižnice* (NSK[105]) u Zagrebu, objavljene od 2015. do 2017. godine. Korpus sadržava 600 000 pojavnica. Korpus je izrađen na način da je dio korpusne građe ručno prikupljen, odnosno znanstveni članci iz područja humanističkih znanosti prikupljeni su ručnom metodom, dok su doktorske disertacije prikupljene automatskom metodom, povlačenjem zadanih mrežnih stranica (v. 4.2.). Iz tako sastavljenoga korpusa dobivene su skice riječi (slika 28) koje pokazuju relacije imenice *pitanje* s drugim vrstama riječi, primjerice glagolima (*postavljati*, *obrađivati*, *istražiti*, *razmatrati*, *iznositi*), odnosno kolokacije specifične za općeznanstveni jezik. Istraživačko pitanje koje se ovdje postavlja jest koje su najčešće glagolsko-imeničke kolokacije iz znanstvenoga polja POSTAVLJANJA PITANJA.

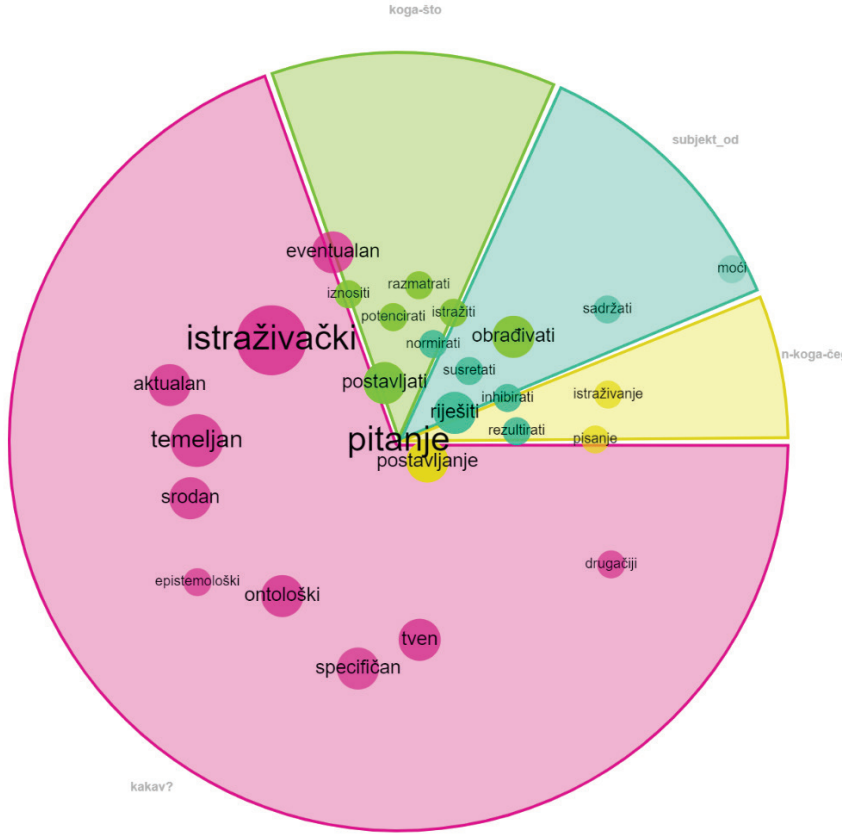
[104] Na korpusu *OZJ* radio je samo jedan istraživač te korpus ima određene nedostatke kao što su duplikati, engleski kontekst i općenito nešto niža kvaliteta teksta (o mrežnim korpusima v. 1.6.).

[105] *Zbirka Disertacija i Magistarskih Radova - Nacionalna i Sveučilišna Knjižnica u Zagrebu* (posjet 12. kolovoza 2019.).

Slika 28: *Skica riječi* pitanje *(korpus OZJ)*

Iz skice riječi prikazane na slici 28 razvidno je da se u općeznanstvenome jeziku glagoli poput *postavljati*, *obrađivati*, *potencirati*, *istražiti*, *razmatrati*, *iznositi* i dr. supojavljuju uz imenicu *pitanje* te čine kolokacije iz znanstvenoga polja postavljanja pitanja. Na temelju dobivenih statističkih podataka iz skica riječi i vizualizacije moguće je izraditi supojavničke profile kolokacija, kao što je to učinjeno za kolokacije općeznanstvenoga njemačkog, engleskog i hrvatskog jezika (Šnjarić i Borucinsky, 2020), a rezultati mogu biti osnovom za sastavljanje jedno- ili višejezičnih (e)rječnika ili glosara koji imaju primjenu u poučavanju jezika te (strojnome i strojno-potpomognutome) prevođenju.

Na slici 29 vizualizirane su relacije glagolsko-imeničkih kolokacija u korpusu *OZJ*.



Slika 29: *Vizualizacija skice riječi* pitanje

Napredne mogućnosti vizualizacije asocijativne snage kolokacije, frekvencije i položaja pruža alat *#LancsBox* (Březina, 2020), a o kolokacijskim grafovima pišu Baker i Egbert (2016), Březina i dr. (2015), te Březina i Pořízka (2021).

7.1.3. ISTRAŽIVANJE SINONIMA

Za pojedina lingvistička istraživanja od manje je važnosti koliko je neka pojava česta, a od veće važnosti koliko je prototipna. To se prije svega odnosi na značenje riječi i njihove semantičke odnose, kao što su sinonimija, antonimija i sl. *SkE*, između ostaloga, podržava opciju pronalaženja sinonima i riječi koje imaju slično značenje na način da su pravila specifična za dani jezik i različita za svaki gramatički odnos (engl. *grammatical relation*, skraćeno

gramrel), a temelji se na distribucijskoj semantici (engl. *distributional semantics*, (Bloomfield, 1993; Firth, 1957; Hall i dr., 1967; Harris, 1954; i dr.), odnosno na jednostavnome načelu da se riječi sličnoga značenja pojavljuju u sličnim kontekstima, pri čemu su manje česte riječi bolje za analizu sinonima.

	Word	Frequency ?	
1	tema	491,636	..
2	stvar	869,313	..
3	problem	1,079,995	..
4	riječ	762,757	..
5	ideja	277,807	..
6	informacija	365,473	..
7	priča	450,281	..
8	situacija	438,987	..
9	odluka	425,145	..
10	slučaj	672,577	..

Slika 30: *Ilustracija funkcionalnosti* Tezaurus *u alat*u SkE

U tablici 12 prikazano je 20 sinonima riječi *pitanje* u tri korpusa hrvatskoga jezika: u korpusu općega jezika (*hrWaC*), korpusu standardnoga jezika (*Riznica*) te specijaliziranome korpusu općeznanstvenoga jezika (*OZJ*) prema učestalosti pojavljivanja.

Tablica 12: *Sinonimi riječi* pitanje *u korpusima* HrWaC, Riznica *i* OZJ

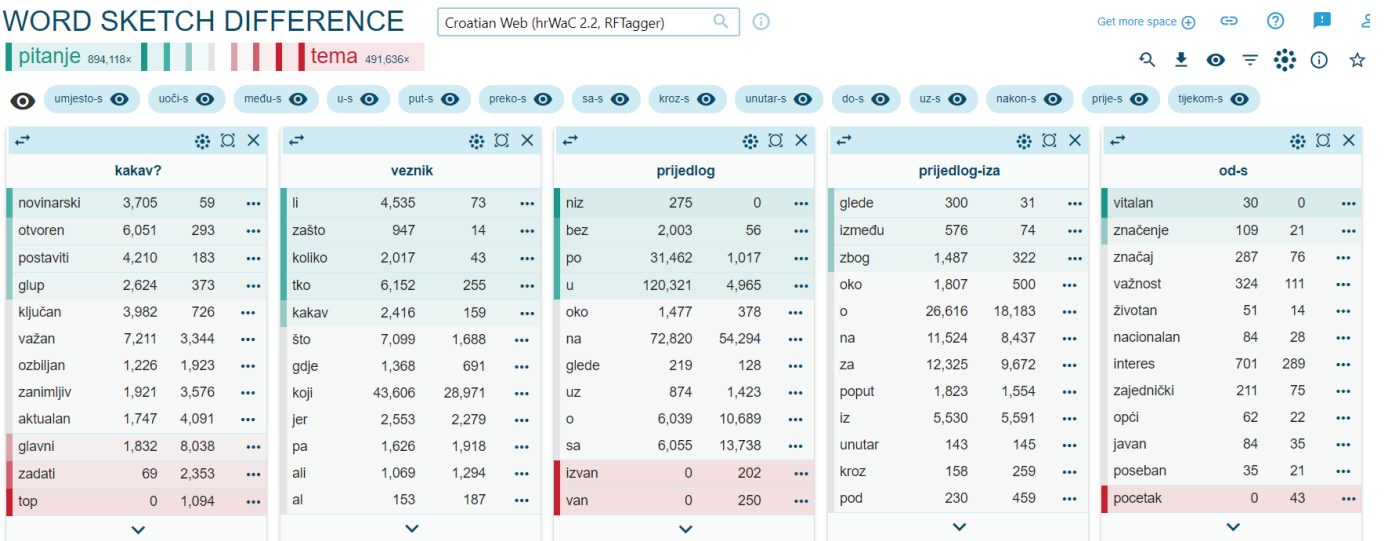
	hrwac	Score 106	Freq	Riznica	Score	Freq	OZJ	Score	Freq
1.	**tema**	0,498	489480	**problem**	0,385	64682	izazov	0,157	36
2.	**stvar**	0,463	865399	**slučaj**	0,31	63410	osnova	0,127	42
3.	**problem**	0,456	1073584	**stvar**	0,304	28206	disciplina	0,12	195
4.	**riječ**	0,432	790266	**tema**	0,303	18351	istra	0,12	48
5.	**ideja**	0,418	277159	**dio**	0,299	92502	metodologija	0,114	70
6.	**priča**	0,417	484977	**odluka**	0,295	47927	zajednica	0,113	61
7.	**informacija**	0,414	364980	posao	0,287	68325	dokument	0,112	117
8.	**situacija**	0,41	437708	mogućnost	0,284	31379	**tema**	0,109	115
9.	**odluka**	0,408	421111	politika	0,284	38209	projekt	0,109	25
10.	**slučaj**	0,406	709670	riječ	0,283	92610	pravilo	0,109	99
11.	način	0,4	937933	prijedlog	0,281	30800	razdoblje	0,109	87
12.	projekt	0,394	619746	projekt	0,281	31631	zemlja	0,108	51
13.	zakon	0,393	561270	program	0,277	46298	literatura	0,104	392
14.	posao	0,391	866471	zakon	0,276	76430	studija	0,102	55
15.	plan	0,391	332746	način	0,273	55211	institucija	0,101	59
16.	mogućnost	0,388	465864	činjenica	0,267	29502	stav	0,101	27
17.	podatak	0,386	458053	zahtjev	0,265	32126	položaj	0,101	31
18.	rješenje	0,384	295050	promjena	0,264	28170	javnost	0,1	13
19.	pravo	0,384	719501	situacija	0,262	26225	**subjekt**	0,099	11
20.	činjenica	0,383	349848	pravo	0,256	63934	promjena	0,098	84
21.	mišljenje	0,382	259302	plan	0,256	26464	smisao	0,098	28
22.	program	0,381	691766	sustav	0,254	36140	faza	0,098	26
23.	odgovor	0,381	290972	odnos	0,254	51357	odluka	0,098	69
24.	politika	0,38	322638	rezultat	0,253	45051	stanje	0,098	50

Sinonimi se automatski detektiraju na temelju konteksta u kojem se pojavljuju. Drugim riječima, riječi koje imaju slične kolokate sličnoga su značenja. Kod skice riječi uspoređuju se sve riječi koje pripadaju istoj vrsti riječi (npr. imenice) s onima s kojima imaju najveći dio zajedničkih kolokata. Rezultat koji se dobije za svaki sinonim pokazuje postotak zajedničkih kolokata. Funkcionalnost *tezaurus* ovisi o bogatim skicama riječi koje sadrže velik broj kolokata, što pak ovisi o velikoj frekvenciji tražene riječi i velikoj frekvenciji

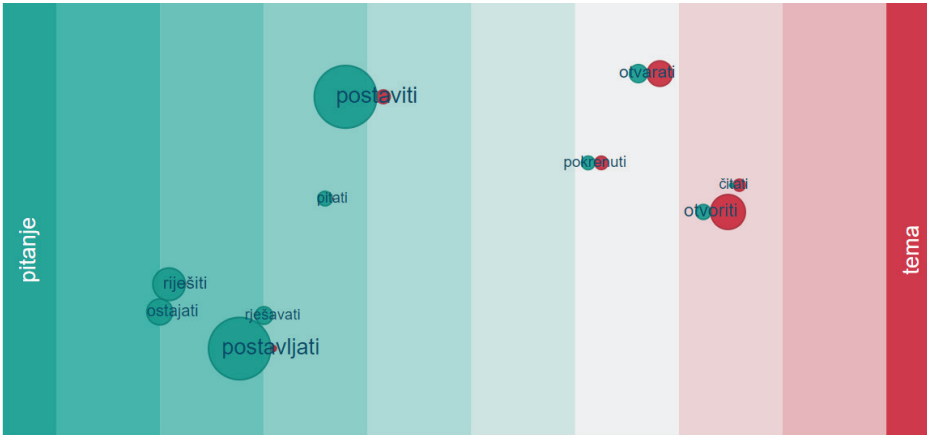
[106] Ovom mjerom izražen je postotak zajedničkih kolokata.

potencijalnih sinonima. Stoga je za ove svrhe potreban jako velik korpus, pri čemu je korpus od 100 000 pojavnica minimalan broj da bi se dobili valjani rezultati. No, tražimo li rjeđe riječi, korpus od više milijardi pojavnica potreban je kako bismo dobili prihvatljiv rezultat. Iz tablice 14 razvidno je da nisu svi rezultati valjani, odnosno da se na popisu nalaze i riječi koje ne dijele semantička obilježja s riječju *pitanje* te nisu njezini sinonimi (npr. *politika*, *institucija*, *javnost* itd.). Masnim slovima u tablici su označene one riječi koje imaju slično značenje s riječju *pitanje*, a iz tablice je također razvidno da je najmanji korpus (*OZJ*) polučio najlošije rezultate. Razlog činjenici da na popisu u tablici 11 ima riječi kojih ne bi trebalo biti leži u automatskome procesuiranju jer se sličnost u značenju ne određuje izravno kao što bi to učinio jezikoslovac, već na temelju usporedbe kolokata. Ako dvije riječi imaju iste kolokate, navest će ih se kao sinonime, iako ne dijele semantička obilježja. Svaki gramatički odnos zasebno se uspoređuje. Također, ako skice riječi nisu pridružene korpusu, ili korpus sadrži drugačije oznake, opcija *tezaurus* koristit će se s univerzalnim skicama riječi s određenim ograničenjima. Podrazumijeva se da korpus koji nije označen i lematiziran neće podržavati ovu funkcionalnost. To, međutim, ne znači da je alat beskoristan, već, kao i u većini jezikoslovnih istraživanja koja se temelje na korpusnim metodama, valja kritički iščitati dobivene rezultate te pronaći uzroke za ovakve pojave.

Na temelju rezultata dobivenih preko tzv. tezaurusa razvidno je da je jedan od bližih sinonima riječi *pitanje* riječ *tema*, posebice u općeznanstvenome jeziku. Usporedimo li skice ovih dvaju riječi pomoću opcije *razlika u skicama riječi*, moći ćemo odrediti po čemu se razlikuju, pri čemu su riječi koje su na slici 31 označene zelenom bojom kolokati imenice *pitanje* (npr. *otvoreno pitanje*), a riječi označene crvenom bojom kolokati imenice *tema* (npr. *glavna tema*), dok su sivom bojom označene one riječi koje se supojavljuju uz obje imenice (npr. *ključno*, *važno*, *aktualno*, *zanimljivo pitanje*; *ključna*, *važna*, *aktualna*, *zanimljiva tema*). Osim kolokacija, iz skica riječi možemo vidjeti i koligacije (npr. *tema/pitanje o*).



Slika 31: *Skice riječi* pitanje *i* tema



Slika 32: *Vizualizacija odnosa* koga-što (glagol + imenica) iz *skica riječi* pitanje i tema

Dakako, korpusni pristup kolokacijama, kao i drugim pojavama, nije savršen, a većina statističkih mjera temelji se na pretpostavci nasumične [107] distribucije što, navodi Kilgarriff (2005), nije slučaj u jezičnoj uporabi, a što je opisano Zipfovim zakonom (v. 3., usp. također Evert, 2005). Dok su kolokacije u engleskome jeziku iscrpno istražene i opisane (Ebeling i dr., 2013; Tognini-Bonelli 2001: 131–156; Xiao i McEnery, 2006; Xiao, 2015), to nije slučaj u hrvatskome jeziku za koji nedostaju kontrastivne studije. No, upravo se takve studije mogu provesti na temelju postojećih korpusa ili korpusa sastavljenih za potrebe određenih istraživanja.

7.1.4. ISTRAŽIVANJE NAZIVLJA

Iduća primjena korpusa koju ćemo ilustrirati jest terminologija. Za potrebe istraživanja jezika (brodostrojarske) struke, autorica je sastavila korpus akademskih tekstova iz područja brodostrojarstva (korpus je nazvan *Brodostrojarstvo*, skraćeno BS) prikupljenih iz časopisa *Naše more* (izdanje 2003. – 2016.), *Pomorstvo* (izdanje 2005. – 2016.) te *Brodogradnja* (2005. – 2017.) dostupnih na portalu znanstvenih časopisa *Hrčak* (usp. Borucinsky i Kegalj, 2017) koji broji oko 1 milijun pojavnica. Usporedimo li ključne riječi iz tablice 13 s ključnim

[107] Kilgarriff (2005: 263) smatra da jezik nije nasumičan, odnosno da govornici nikada nasumce ne biraju jezične jedinice.

riječima iz tablice 7, razvidno je da su najčešće riječi u ovome korpusu (*motor*, *koljenast*, *porivni*) riječi koje nećemo naći u općem leksiku.

Tablica 13: *Najfrekventnije riječi, POS oznake u specijaliziranome korpusu jezika brodostrojarske struke*

Red. Broj.	Top 5 ključnih riječi (različnica)	Top 5 vrsta riječi
1.	motor	Mrc
2.	koljenast	Mdm
3.	porivni	Npnsn
4.	klipni	Mic
5.	klip	Np

Za specijalizirane korpuse najkorisnije su funkcionalnosti *N-gram* i *crpljenje nazivlja* (v. 3.). Potonja, koja se također naziva i ključnost identificirat će sve riječi (pojavnice), neovisno o tome kojoj vrsti riječi pripadaju, iako se u pravilu detektiraju samo imenice i pridjevi jer su frekvencije drugih vrsta riječi usporedive u tekstovima. Analiza frekvencija korisna je za ispitivanje pojava onih riječi za koje očekujemo da će nam dati značajne podatke. No, ako nemamo očekivanja o tome koje će riječi biti informativne, možemo preko analize ključnih riječi otkriti statistički značajne riječi u korpusu. Potom je najbolje provesti analizu konkordancija kako bismo pomnije ispitali ključne riječi dobivene u korpusu.

Funkcionalnost *N-gram* proizvodi frekvencijske liste niza pojavnica. Korisnik pri radu s ovom funkcionalnošću koja se još naziva i dohvaćanje podataka (engl. *retrieval*), određuje broj N-grama koje želi izlistati te minimalnu frekvenciju kojom se pojavljuju u korpusu, koja se u pravilu postavlja na 10 do 40 pojava na milijun riječi (Biber, 2006). Nizovi od četiri riječi, tj. kvadrigrami, smatraju se proširenim kolokacijskim svezama i prilično su česti, nizovi od tri riječi nešto se rjeđe istražuju jer su u pravilu sadržani u kvadrigramima, dok su nizovi od pet ili više riječi frazalne prirode i manje frekventni (Hyland, 2008). Stoga se većina istraživanja ove vrste višerječnih naziva provodi na kvadrigramima. Korisnik i pomoću regularnih izraza može specificirati za koje N-grame želi dobiti frekvenciju iz korpusa. Rezultati se mogu dobiti za kombinaciju riječi te minimalnu frekvenciju, a u tablici 14 prikazani su N-grami iz korpusa pomorsko-pravnih tekstova na hrvatskome jeziku, s apsolutnom i relativnom frekvencijom u korpusu (stupac 2 i 3), apsolutnom i relativnom frekvencijom u dokumentu (stupci 4 i 5) te srednjom reduciranom frekvencijom (v. 3.).

Tablica 14: *Kvadrigrami iz korpusa pomorsko-pravnih tekstova*[108]

Kvadrigram	AF	RF	DOCF	Relativna DOCF	ARF
“a ako je brod”	11	1.038.467	5	3.846.154	536.159
“ako je brod u”	21	1.982.527	5	3.846.154	717.430
“ako je tanker za”	11	1.038.467	5	3.846.154	525.069
“ako je to moguće”	13	1.227.279	7	5.384.615	912.490
“ako je to potrebno”	19	1.793.715	8	6.153.846	1.139.256
“ako postoji osnovana sumnja”	23	2.171.339	6	4.615.385	930.317
“ako se brod nalazi”	15	1.416.091	6	4.615.385	653.554
“ako se radi o”	24	2.265.746	7	5.384.615	1.044.457
“ako se u slučaju”	11	1.038.467	5	3.846.154	681.416
“ako to nije moguće”	10	944.061	5	3.846.154	707.964
“bez obzira je li”	12	1.132.873	6	4.615.385	796.697
“bez obzira na odredbe”	30	2.832.182	7	5.384.615	1.165.553
“bi se osiguralo da”	58	5.475.552	6	4.615.385	2.207.890
“biti u skladu s”	47	4.437.085	8	6.153.846	2.149.495
“čiju zastavu brod vije”	19	1.793.715	6	4.615.385	1.156.643
“će se da je”	31	2.926.588	6	4.615.385	1.489.276
“će se da se”	14	1.321.685	5	3.846.154	600.763
“će se izdati samo”	14	1.321.685	5	3.846.154	546.952
“će se primijeniti na”	16	1.510.497	5	3.846.154	620.736
“će stupiti na snagu”	19	1.793.715	5	3.846.154	813.484
“će šteta vjerojatno nastati”	15	1.416.091	5	3.846.154	589.161
“da bi se utvrdilo”	11	1.038.467	5	3.846.154	495.595
“da brod nije sposoban”	11	1.038.467	5	3.846.154	548.399
“da brod u svakom”	20	1.888.121	5	3.846.154	657.759
“da će šteta vjerojatno”	15	1.416.091	5	3.846.154	589.161
“da je brod u”	13	1.227.279	6	4.615.385	601.090
“dok ne bude mogao”	13	1.227.279	5	3.846.154	535.673
“dužan je o tome”	15	1.416.091	6	4.615.385	819.294

[108] Usporedni korpus pomorskopravnih tekstova sastavljen je za potrebe istraživanja leksičkih obrazaca u ovome specifičnom žanru, a rezultati tog istraživanja objavljeni su u časopisu *ICAME* (Borucinsky i Pritchard, 2022).

“dužna je osigurati da”	10	944.061	6	4.615.385	607.694
“ima pravo na naknadu”	13	1.227.279	5	3.846.154	666.094
“je izmijenjen i dopunjen”	11	1.038.467	6	4.615.385	477.021
“je izmijenjena i dopunjena”	24	2.265.746	7	5.384.615	1.057.681
“je međunarodna konvencija o”	12	1.132.873	7	5.384.615	577.172
“je o tome obavijestiti”	11	1.038.467	6	4.615.385	688.781
“je sklopljen ugovor o”	19	1.793.715	5	3.846.154	545.686
“je to prikladno i”	21	1.982.527	5	3.846.154	729.219
“je u skladu s”	25	2.360.152	8	6.153.846	1.440.775
“jedne godine od dana”	13	1.227.279	5	3.846.154	542.490
“kad je to potrebno”	13	1.227.279	5	3.846.154	747.682
“kada je brod u”	10	944.061	6	4.615.385	565.322
“kada je to potrebno”	11	1.038.467	7	5.384.615	577.639
“kada se smatra da”	10	944.061	5	3.846.154	527.674
“kako bi se osiguralo”	66	6.230.800	7	5.384.615	2.646.311
“kako je izmijenjena i”	26	2.454.558	7	5.384.615	1.079.265
“kako je navedeno u”	88	8.307.734	9	6.923.077	3.200.693
“kako je utvrđeno u”	72	6.797.237	5	3.846.154	1.956.652
“kako su navedene u”	10	944.061	5	3.846.154	470.299
“koja je navedena u”	10	944.061	5	3.846.154	459.151
“koja se odnose na”	40	3.776.243	9	6.923.077	2.046.211

Učestali leksički spojevi funkcioniraju kao jedna gramatička cjelina, iako često prelaze granice gramatičkih kategorija (npr. *je u skladu s,* će se primijeniti *na* itd.), a strukturno ih se može podijeliti na imenske, glagolske, priložne, surečenične te pridjevske. Prema njihovoj funkciji učestali leksički spojevi mogu biti orijentirani k istraživanju, sudioniku ili intertekstualni. S obzirom na to da za hrvatski jezik ne postoje istraživanja o učestalim leksičkim spojevima, ova funkcionalnost može dati značajan doprinos opisu žanra i funkcionalnih stilova, kao i pokazati potrebu za ciljnim podučavanjem formulaičnih izraza u stranome jeziku struke koji utječu na jezičnu produkciju neizvornih govornika (Borucinsky i Pritchard, 2022).

Zaključno, korpusni podaci pružaju pouzdane empirijske dokaze za identifikaciju leksičkih obrazaca i njihovih odnosa, a tako dobivene spoznaje mogu

poboljšati opis jezika općenito, ali i dati uvid u sitne razlike među riječima koje imaju slično značenje. Dobiveni podaci u konačnici imaju konkretnu primjenu u nastavi jezika (struke), prevođenju, leksikografiji, terminografiji, kontrastivnoj lingvistici i drugim jezikoslovnim (pod)disciplinama. No, mogu dati i poticaj za postavljanje novih ili boljih opisa jezičnih struktura.

7.2. KORPUSNO ISTRAŽIVANJE GRAMATIKE

Proučavanjem gramatike i leksikogramatike u korpusu dobit ćemo uvid u to kako riječi, skupine, (su)rečenice i iskazi oblikuju značenje u danome ko(n) tekstu. Na temelju podataka iz korpusa možemo potvrditi ili opovrgnuti intuitivne pretpostavke o gramatičkim obrascima u jeziku (v. pristup utemeljen na korpusu, poglavlje 1.2). Halliday i dr. (2014) navode da korpus također omogućuje pristup autentičnim tekstovima te kvalitativnu i kvantitativnu analizu gramatičkih obrazaca u kontekstu. Tako, primjerice, možemo utvrditi da je nominalizacija[109] tipičnija u jeziku struke u odnosu na opći jezik. Kegalj (u pripremi) pokazala je da pomorskopravne tekstove karakterizira tendencija ka nominalizaciji koja se očituje u većemu broju imenica u odnosu na glagole, osobito u većoj frekvenciji glagolskih imenica i naglašenoj tendenciji uporabe perifrastičnih glagola koji otvaraju mjesto imenskoj skupini kao njihovoj semantičkoj dopuni. Nadalje, Borucinsky i Kegalj (2019) pokazale su kako pretraživati glave imenske skupine u korpusu kako bi se ispitala sintaktička višeznačnost u jeziku struke. Korpusne metode mogu dati odgovor na pitanje je li pisani jezik sintaktički složeniji od govorenoga na način da se mjere prosječna duljina komunikacijske jedinice kao jedna od mjera sintaktičke složenosti, te provjeri koje sintaktičke strukture su zastupljenije u pisanome u odnosu na govoreni jezik ili da se ispita frekvencija diskursnih oznaka u korpusu pisanoga i govorenoga jezika. Takvo istraživanje proveli su Hržica i dr. (2021).

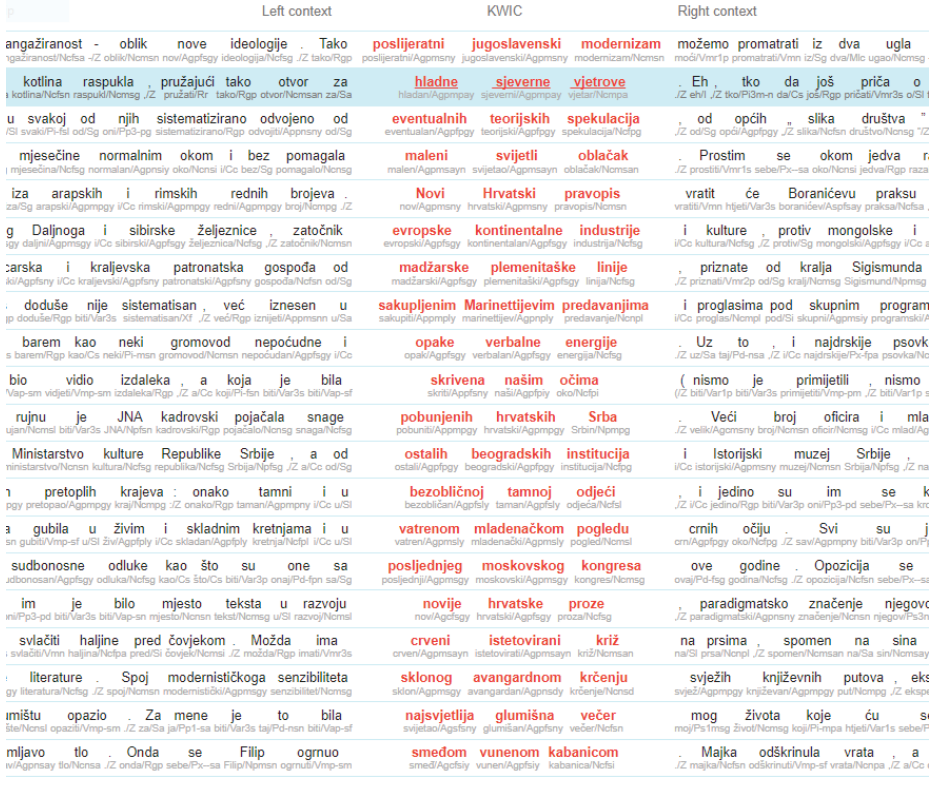
7.2.1. IMENSKA SKUPINA

Sintaktička razina predstavlja najveći izazov korpusnim metodama jer se sintaktičke strukture ne mogu pretraživati automatski, već samo prema vrstama riječi. Primjerice, ako promatramo imensku skupinu koja ima strukturu A+A+N (npr. *najčešće postavljana pitanja*, *lijep sunčan dan*, žive jezikoslovne

[109] Nominalizacija je izražavanje određene pojavnosti imenskom skupinom umjesto surečenicom (Borucinsky 2015: 204).

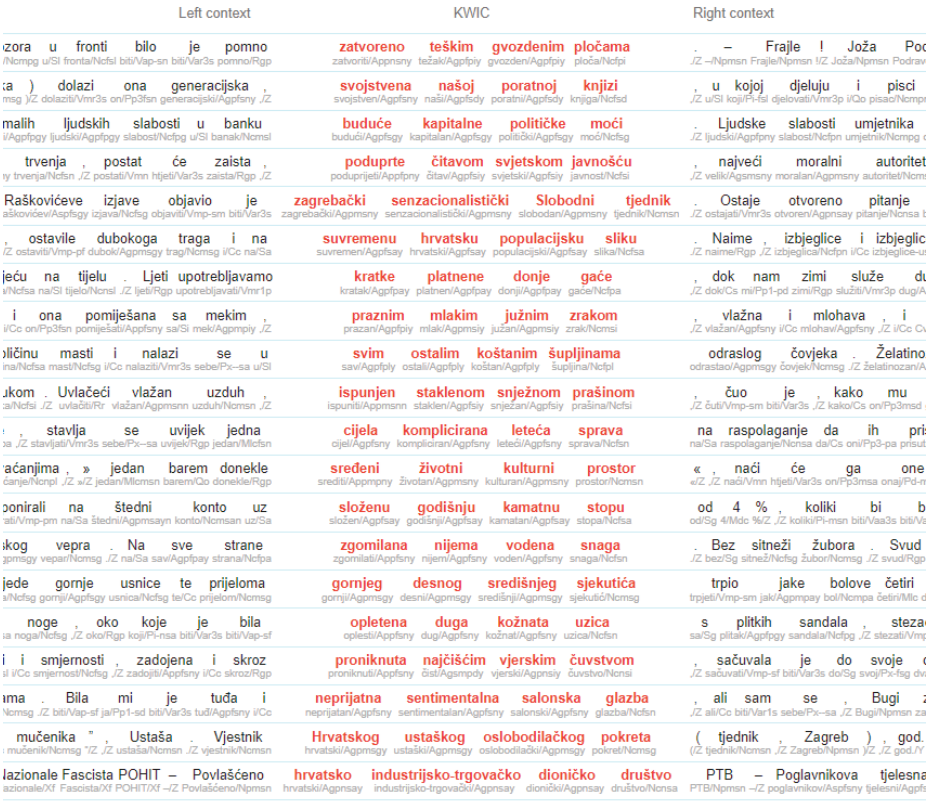
rasprave, *lijepa naša Hrvatska* itd.), odnosno imensku skupinu u kojoj je glava skupine imenica, a modificiraju je dva pridjeva, ne možemo računalu dati naredbu da u tekstu pronađe sve imenske skupine ili sintagme (NP), već moramo postaviti ciljanu pretragu preko vrsta riječi (npr. [tag="A.*"] [tag="A.*"] [tag="N.*"]). Parseri i univerzalne ovisnosti daju opis i omogućuju pretragu sintaktičkih struktura, no i dalje ostaje neriješen problem definicije imenske skupine u hrvatskome, npr. jesu li determinatori poput *ovaj*, *onaj* itd. dijelom imenske skupine ili zasebne kategoije, kako se tretiraju postmodifikatori i sl.

Na slici 33 prikazan je konkordancijski niz takve pretrage u korpusu *Riznica*.

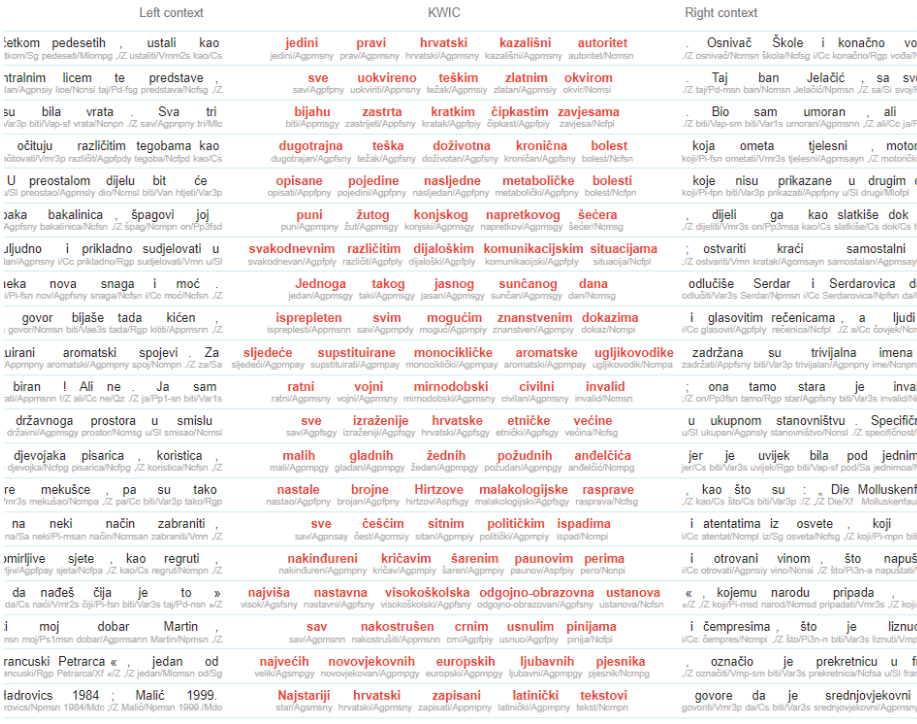


Slika 33: *Konkordancijski niz pretrage* [tag="A.*"] [tag="A.*"] [tag="N.*"] *u* Riznici, *slučajni uzorak*

Međutim, imenska skupina u pravilu je znatno složenija te može imati i strukturu A+A+A+N (slika 34) ili A+A+A+A+N (slika 35).



Slika 34: *Konkordancijski niz pretrage* [tag="A.*"] {3} [tag="N.*"] *u* Riznici, *slučajni uzorak*



Slika 35: *Konkordancijski niz pretrage* [tag="A.*"] {4} [tag="N.*"] *u* Riznici, *slučajni uzorak*

Iz korpusa, kao što je već navedeno, možemo dobiti podatke o evidenciji (postoji li u hrvatskome imenska skupina u kojoj pet ili više pridjeva premodificira imenicu) i frekvenciji (je li češća imenska skupina s dva, tri, četiri ili pet pridjeva ispred imenice), a konkordancijski prikazi korisni su i za proučavanje redoslijeda kojim pridjevi modificiraju imenicu, i zašto je, primjerice, struktura *sređen životni kulturni prostor* gramatički prihvatljivija u odnosu na strukturu *životni kulturni sređen prostor. Odgovor na ovo pitanje dala je Borucinsky (2015), a on leži u suodnosu sintakse i semantike, odnosno vrsta modifikatora koji se nalaze bliže ili dalje od glave imenske skupine. Primjerice, klasifikatori poput pridjeva *kulturni* imaju najuže značenje i zbog toga su najbliže glavi skupine. Što pridjev ima šire značenje, bit će udaljeniji od glave imenske skupine. Ova se pretpostavka može potvrditi i pretragom korpusa prikazanom u tablici 15, u kojoj je prikazan konkordancijski niz u kojemu ključne riječi jesu tri pridjeva od koji je jedan *jedinstveni*.

Tablica 15: *Konkordancijski niz pretrage* [tag="A.*"]{3 containing[word="jedinstveni"]

1.	i ekološki prihvatljiva rješenja za potrebe uredskih poslova. Prednosti Kyocerina dokazana upravljačka jedinica i	**jedinstveni/Agpmsny inteligentni/Agpmsny programski/Agpmsny**	jezik PRESCRIBE, sada se imprementiraju u sve Kyocera Mita pisače i multifunkcijske uređaje (MFP). Korištenjem
2.	razvoja grada analogno je onovremenim europskim uzorima. Zaslugom Viktora Axmanna grad Osijek se ulančao u	**jedinstveni/Agpmsayn urbarhitektonski/ Agpmsayn hrvatski/ Agpmsayn**	i europski korpus moderne 20. stoljeća zaključio jer dr. Ambruš. Inače, Viktor Axmann je u Osijeku projektirao više od
3.	mentalnog pokreta, tekućeg procesa preobrazbi. Ukratko, izložba bi u ‘ malom uzorku ‘ trebala sublimirati	**jedinstveni/Agpmsayn Kožarićev/Aspmsnn dinamični/Agpmsny**	vitalizam koji ga čini mladim i u njegovim objektivno poznim godinama. U tom smislu, istaknuti motiv falusa koji autor
4.	je u Dubrovniku u čuvenoj Vila Šeherezada uz čiju se povijest... U kreativnoj i opuštenoj atmosferi, jučer je svečano	**otvoren/Appmsnn jedinstveni/Agpmsny modni/Agpmsny**	prostor u Zagrebu zvan vintage boutique “ Oko “. U staroj jezgri Zagreba, na adresi Opatovina 13 nalazi se ovaj poseban
5.	vratiti koži njezinu svježinu, luminioznost i prirodan izgled, jer djeluje trostruko: - posvjetljuje i pruža jedan	**kompletni/Agpmsny jedinstveni/Agpmsny novi/Agpmsny**	tonus koži - sprječava produkciju melanina te uklanja mrlje i pjege. Za ove rezultate treba duži period da postanu
6.	sve veći i veći, implodiraju i stvaraju veliku količinu energije koja potiče raspadanje membrana masnih stanica. Med	**Contourov/Aspmsnn jedinstveni/Agpmsny ultrazvučni/Agpmsny**	tretman razbija membrane masnih stanica, masne stanica se uništavaju, ispuštajući masne kiseline i trigliceride.
7.	“ u dnevnoj potrošnji elektroenergetskog sistema. Budući rad termoelektrane moguće je sagledati jedino kroz	**jedinstveni/Agpmsayn integralni/Agpmsayn elektroenergetski/ Agpmsayn**	sustav Hrvatske. Glavni pravci razvoja usmjereni su na revitalizaciju i dogradnju kapaciteta. U tu svrhu unutar
8.	u Crikvenici, turistička zajednica je objavila natječaj za Pavlinsku tortu. Na taj će način Crikvenica uskoro dobiti	**novi/Agpmsayn jedinstveni/Agpmsayn gastronomski/Agpmsayn**	suvenir Pavlini su, sve do ukinuća Najam kampera Branko Ivanković i NK Dinamo sporazumno su raskinuli međusobni
9.	revolucionarni hibridni sustav u Optimu koja se može pohvaliti brojnim međunarodnim dizajnerskim nagradama. Zato	**jedinstveni/Agpmpny sportski/Agpmpny lijevani/Agpmpny**	naplaci, spušteno podvozje i dorađena karoserijska aerodinamika predstavljaju samo dio govora tijela najnovijeg
10.	raznolikošću i književno-jezičnom poviješću hrvatski je jezik samosvojna pojavnost iz koje je izrastao i	**jedinstveni/Agpmsayn suvremeni/Agpmsayn standardni/Agpmsayn**	jezik. Briga za hrvatski jezik bit će pojačana donošenjem zakona o njegovoj službenoj uporabi, s podzakonskim aktima

11.	vodilo sjedinjenju tih dviju crkava, a u daljnjem tijeku vrlo vjerojatno sjedinjenju hrvatskoga i srpskoga naroda u	**jedinstveni/Agpmsayn umjetni/Agpmsayn jugoslavenski/ Agpmsayn**	narod, odnosno u stvari u velikosrpski narod. Titova ponuda Kad u tome komunisti nisu uspjeli uslijedilo je prvo
12.	kao i sve većim i sigurnijim crtačkim umijećem. U ciljniku crtačeva pera duhovna je meta njegov likovni motiv grad, taj	**jedinstveni/Agpmsayn južnohrvatski/Agpmsayn urbanistički/Agpmsayn**	pleter u povijesnoj jezgri romaničkog, gotičkog ili renesansnog sloga, u skladnom rimovanju horizontale trga i
13.	Tijelovo je zapovjedna svetkovina i raspored svetih Misa u Katedrali je kao i nedjeljom. U podmorju poluotoka Vižula	**pronađen/Appmsnn jedinstveni/Agpmsny rimski/Agpmsny**	amulet u obliku falusa Rimski amulet koji je bogatim vlasnicama služio kao zaštita protiv neplodnosti pronađen je
14.	Presvetog Otkupitelja. Ujedinjavanjem Teologije u Splitu i Franjevačke visoke teološke škole iz Makarske u	**jedinstveni/Agpmsayn Katolički/Agpmsayn bogoslovni/Agpmsayn**	fakultet Sveučilišta u Splitu, franjevački bogoslovi sa svojim odgojiteljima i profesorima preselili su u Split.
15.	ovaj automobil čine jedinstvenim, Zonda 750 najprepoznatljivija je po neobičnoj boji. Zonda 750 nije prvi, a možda ni	**posljednji/Agpmsny jedinstveni/Agpmsny oproštajni/Agpmsny**	model. Prije toga je softverski mogul David Heinemeier Hansson kupio jedinstvenu Zondu HH, a katarska kraljevska
16.	organizma, koji smetaju začeću i ukoliko se ustvrdi njihova prisutnost, potrebno ih je odgovarajuće liječiti.	**Pokrenut/Appmsnn jedinstveni/Agpmsny mrežni/Agpmsny**	priručnik o turističkoj kulturi 05.04.2013 Ovogodišnje izdanje obrazovnoga projekta Turistička kultura
17.	Europske komisije za pravosuđe Jacques Barrot izrazio je žaljenje što je u samo 10 zemalja članica EU-a	**operativan/Agpmsnn jedinstveni/Agpmsny telefonski/Agpmsny**	broj 116000 na koji se može prijaviti nestanak djeteta te zatražio da se što prije otkloni taj nedostatak. ‘
18.	Compagnie u Marseillesu, Francuska. U Zagrebačkom kazalištu lutaka u petak, 14. lipnja, u 20 sati, premijerno će biti	**izveden/Appmsnn jedinstveni/Agpmsny glazbeno-scenski/ Agpmsny**	projekt ‘ Now What? ‘ koji okuplja niz kreativaca iz područja glazbe, videoumjetnosti i scenske umjetnosti te
19.	i Valter Lacman. NY Vice Verses “ ZG Edition - poetsko glazbeni rusvaj “ U četvrtak, 21.2. u 20 sati u Booksi će se održati	**jedinstveni/Agpmsny međunarodni/Agpmsny glazbeno-poetski/ Agpmsny**	događaj naziva Vice Verses: Zagreb Edition. Radi se o ciklusu slam poetry i glazbenih večeri koje je u SAD-u pokrenula
20.	55 u Lisinskom te potom na gostovanjima u Puli i Rijeci pokazali su da nisu potr... 26. travnja 2012. godine MONTAЖ $	**TROJ-ev/Aspmsnn jedinstveni/Agpmsny kulturni/Agpmsny**	događaj 55 ostvario je svoje prvo gostovanje u Istarskom narodnom kazalištu u Puli. Protagonisti kulturnog događaja

Iz navedenih primjera možemo zaključiti da pridjev *jedinstveni* kao premodifikator glave skupine, tj. imenice, zauzima položaj koji je udaljeniji od glave skupine, nalazi se dakle na položaju koji nije neposredno uz glavu skupine (pod uvjetom da nije jedini pridjev koji premodificira imenicu).

Ovakvi podaci vrijedan su izvor informacija za razumijevanje jezičnih pojava, u ovome slučaju imenske skupine, te u konačnici vode ka boljemu teorijskom opisu, posebice iz razloga što status imenske skupine, fraze ili sintagme nije riješen u hrvatskim gramatikama. Pod time se misli da većina hrvatskih gramatika naglasak stavlja na strukturu surečenice i rečenice, dok je struktura skupine ili sintagme zanemarena, a u hrvatskome, za razliku od engleskoga jezika (npr. Biber i dr., 1999; Carter i McCarthy, 2006), nema korpusno utemeljene gramatičke tradicije (v. također Borucinsky, 2015; Borucinsky, 2017). Kao što je opisano u poglavlju 1.2., jedan od načina na koji se računalnojezikoslovni resursi i alati u jezikoslovnim istraživanjima rabe jest da se u korpusu traže unaprijed zadani jezični obrasci. U većini slučajeva to su leksički obrasci (v. 4.1.), no korpusi, kao što je pokazala Borucinsky (2017), mogu poslužiti i za utvrđivanje sintaktičkih obrazaca i kao osnova za oblikovanje sintaktičke teorije ili pisanje gramatika. Autorica (ibid.) također je pokazala kako se sintaktička obilježja flektivnih, odnosno morfološki bogatih jezika, mogu promatrati u korpusu.

U ovome dijelu osvrnut ćemo se detaljnije na problematiku imenske skupine. Pojednostavljeno, imenska skupina sastoji se od determinatora i premodifikatora (riječi koje se nalaze lijevo od glave skupine), glave skupine (u pravilu imenice) i postmodifikatora (riječi koje se nalaze desno od glave skupine), ili:

IMENSKA SKUPINA

DETERMINATOR *PREMODIFIKATOR* *GLAVA* *POSTMODIFIKATOR*

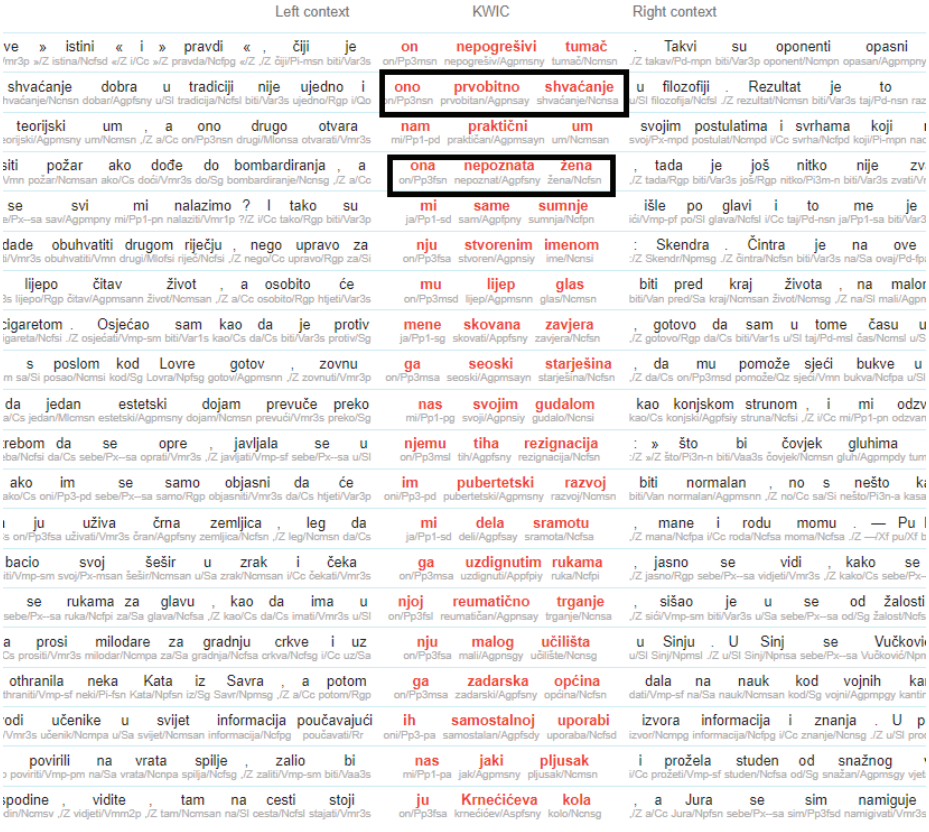
Pomoću upita preko CQL-a i regularnih izraza (v. 3.) iz korpusa se na jednostavan način mogu dobiti popisi imenica u određenome padežu (tablica 16). Velika prednost jezično anotiranoga korpusa očituje se u brzini i količini jezičnih podataka koje možemo dobiti, posebice u usporedbi s ručnom pretragom koja bi bila dugotrajna, zamorna, a na koncu bi i rezultat ručne pretrage jezikoslovaca mogao biti upitan s obzirom na mogućnost pogreške koja se istraživačima može potkrasti.

Također, možemo utvrditi i redoslijed pojavljivanja determinatora i modifikatora unutar imenske skupine, odnosno je li uobičajenija imenska skupina *taj njegov/njezin zahtjev* [lemma="taj"][lemma="njegov|njezin"][tag="N.*"] ili *njegov taj zahtjev* [lemma="njegov|njezin"][lemma="taj"][tag="N.*"]. Prva pretraga dala je 993 rezultata, a druga 1 rezultat u *Riznici*. Evidencija će pokazati i je li takav poredak uvijek isti za pridjeve i kojim se redoslijedom pokazne zamjenice javljaju u korpusu.

Tablica 16: *Redoslijed zamjenica ispred imenice*

Pretraga	**RF hrWac**	**RF Riznica**	**RF HNK (2015)**	**Primjer**
pokazni determinativ + posvojni determinativ + imenica [tag="Pd.*"] [tag="Ps.* “] [tag="N.*"]	70,29	41,92	20,2	taj njegov zahtjev, ta moja zamisao
posvojni determinativ + pokazni determinativ + imenica [tag="Ps.*"] [tag="Pd.*"] [tag="N.*"]	0,62	0,3	0,1	njegov takav pokušaj; njihovo ovakvo ponašanje
neodređeni determinativ + posvojni determinativ + imenica [tag="Pi.*"] [tag="Ps.*"] [tag="N.*"]	42,4	33,43	22,9	nekih naših ljudi; svaki njezin nastup
posvojni determinativ + neodređeni determinativ + imenica [tag="Ps.*"] [tag="Pi.*"] [tag="N.*"]	0,76	0,35	0,2	njegov svaki istup; moje nekakvo iskustvo

Ovi rezultati, iako ne apsolutno točni, mogu pokazati tendencije u korpusu te biti osnovom za daljnja kvantitativna promatranja i kritička promišljanja. No, pretrage sintaktičkih obrazaca u korpusu od korisnika zahtijevaju poseban oprez, što je prikazano u sljedećim primjerima. Pretpostavimo da posvojne zamjenice stoje ispred pridjeva koji modificiraju imenice (npr. *moja nova torba*), za što bismo mogli postaviti pretragu [tag="Pp.*"] [tag="A.*"] [tag="N.*"], no također pretpostavimo da je moguće i obrnuto: da pridjevi stoje ispred posvojnih zamjenica iza kojih slijedi imenica (npr. ?*nova moja torba*), za što bismo postavili pretragu [tag="A.*"] [tag="Pp.*"] [tag="N.*"]. Međutim, pretrage neće dati očekivani rezultat jer postavljeni obrazac pretrage nadilazi granice imenske skupine, što je razvidno iz 18 od 20 primjera prikazanih na slici 36. Zadana pretraga dala je samo dva željena rezultata (*ono prvobitno shvaćanje, ona nepoznata žena*). Na slici 36 vidljivo je da nijedan od dobivenih rezultata ne odgovara očekivanome obrascu.



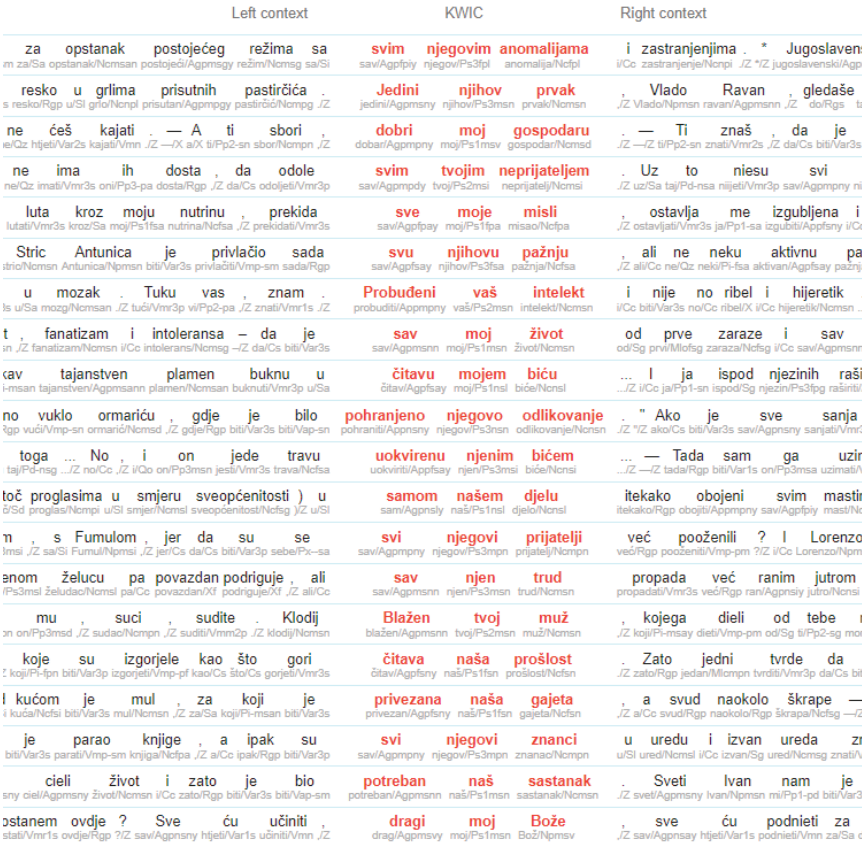
Slika 36: *Konkordancijski niz pretrage* [tag="Pp.*"] [tag="A.*"] [tag="N.*"] *u* Riznici, *slučajni uzorak*

Slika 37: *Konkordancijski niz pretrage* [tag="A.*"] [tag="Pp.*"] [tag="N.*"] *u* Riznici, *slučajni uzorak*

Međutim, ne treba dobivene podatke odmah odbaciti, već se njihovim iščitavanjem mogu uočiti obrasci koje nismo tražili u korpusu (npr. *zračne mi stope*, *rodne im grude*, *lijepi im ćaća*), a koji mogu biti polazištem za daljnja ili druga jezikoslovna propitivanja, poput onoga je li struktura „pridjev + osobna zamjenica + imenica“ karakterističnija za razgovorni stil.

Nešto bolje rezultate dobit ćemo za pretrage [tag ="Ps.*"] [tag ="A.*"] [tag="N.*"] (slika 38) i [tag ="A.*"] [tag ="Ps.*"] [tag="N.*"] (slika 39).

Slika 38: *Konkordancijski niz pretrage* [tag ="Ps.*"] [tag ="A.*"] [tag="N.*"] *u* Riznici, *slučajni uzorak*



Slika 39: *Konkordancijski niz pretrage* [tag ="A.*"] [tag ="Ps.*"] [tag="N.*"] *u* Riznici, *slučajni uzorak*

Svakako bi valjalo i pomnije istražiti pogreške u morfosintaktičkome označavanju onoga dijela korpusa koji se odnosi na sastavnice unutar imenske skupine te utvrditi njihove uzroke, čime bi se pomoću razjašnjenih kriterija o jedinicama unutar imenske skupine, kojima pripadaju i modifikatori, znatno preciznije mogao označiti korpus, što bi imalo utjecaja na buduća računalnojezikoslovna istraživanja o hrvatskome jeziku. U nastavku knjige navodi se nekoliko primjera kako bi se ilustrirala ta problematika.

U tablici 17 prikazano je prvih 20 rezultata pretrage [tag="N.*g.*"][110] u *hrWaC*-u.

[110] Imenica u hrvatskome ima pet atributa, a ova pretraga znači: „Pronađi sve imenice (opće i vlastite), bilo kojega roda, u jednini ili množini, u genitivu, žive i nežive.“ O regularnim izrazima v. 2.2.

Tablica 17: *Konkordancijski niz pretrage* [tag="N.*g.*"] *u* hrWaC-*u*

1.	Uštedite korištenjem	**plina/Ncmsg**	... Prirodni plin upotrebljava se u kućanstvu za grijanje prostora, pripremu tople vode, termičku obradu hrane, a
2.	Uštedite korištenjem plina... Prirodni plin upotrebljava se u kućanstvu za grijanje	**prostora/Ncmsg**	, pripremu tople vode, termičku obradu hrane, a koristi se i sve više za hlađenje prostora. Prirodni plin nalazi svoju
3.	Uštedite korištenjem plina... Prirodni plin upotrebljava se u kućanstvu za grijanje prostora, pripremu tople	**vode/Ncfsg**	, termičku obradu hrane, a koristi se i sve više za hlađenje prostora. Prirodni plin nalazi svoju široku potrošnju u
4.	plina... Prirodni plin upotrebljava se u kućanstvu za grijanje prostora, pripremu tople vode, termičku obradu	**hrane/Ncfsg**	, a koristi se i sve više za hlađenje prostora. Prirodni plin nalazi svoju široku potrošnju u industriji za tehnološke
5.	se u kućanstvu za grijanje prostora, pripremu tople vode, termičku obradu hrane, a koristi se i sve više za hlađenje	**prostora/Ncmsg**	. Prirodni plin nalazi svoju široku potrošnju u industriji za tehnološke procese, kao gorivo ili kao sirovina, te za
6.	svoju široku potrošnju u industriji za tehnološke procese, kao gorivo ili kao sirovina, te za grijanje i hlađenje	**prostora/Ncmsg**	. Kontaktirajte nas... ZELENJAK PLIN d. o. o. ovlašten je od HRVATSKE STRUČNE UDRUGE ZA PLIN za ispitivanje plinskih
7.	, te za grijanje i hlađenje prostora. Kontaktirajte nas... ZELENJAK PLIN d. o. o. ovlašten je od HRVATSKE STRUČNE	**UDRUGE/Ncfsg**	ZA PLIN za ispitivanje plinskih instalacija. Distribucija prirodnog plina obavlja se na području Grada Klanjca, te
8.	. Kontaktirajte nas... ZELENJAK PLIN d. o. o. ovlašten je od HRVATSKE STRUČNE UDRUGE ZA PLIN za ispitivanje plinskih	**instalacija/Ncfpg**	. Distribucija prirodnog plina obavlja se na području Grada Klanjca, te općina Tuhelj, Kumrovec, Zagorska sela,
9.	d. o. o. ovlašten je od HRVATSKE STRUČNE UDRUGE ZA PLIN za ispitivanje plinskih instalacija. Distribucija prirodnog	**plina/Ncmsg**	obavlja se na području Grada Klanjca, te općina Tuhelj, Kumrovec, Zagorska sela, Kraljevec na Sutli, te manji dio na
10.	STRUČNE UDRUGE ZA PLIN za ispitivanje plinskih instalacija. Distribucija prirodnog plina obavlja se na području	**Grada/Ncmsg**	Klanjca, te općina Tuhelj, Kumrovec, Zagorska sela, Kraljevec na Sutli, te manji dio na području općina Desinić,
11.	UDRUGE ZA PLIN za ispitivanje plinskih instalacija. Distribucija prirodnog plina obavlja se na području Grada	**Klanjca/Npmsg**	, te općina Tuhelj, Kumrovec, Zagorska sela, Kraljevec na Sutli, te manji dio na području općina Desinić, Veliko

12.	se na području Grada Klanjca, te općina Tuhelj, Kumrovec, Zagorska sela, Kraljevec na Sutli, te manji dio na području	**općina/Ncfpg**	Desinić, Veliko trgovišće i Dubravice. Natječaj - Tipski dječji vrtić drvene konstrukcije Komunalna naknada
13.	Tuhelj, Kumrovec, Zagorska sela, Kraljevec na Sutli, te manji dio na području općina Desinić, Veliko trgovišće i	**Dubravice/Npfsg**	. Natječaj - Tipski dječji vrtić drvene konstrukcije Komunalna naknada Djelatno vrijeme: 08:00 15:00 Pauza: 11:00
14.	na Sutli, te manji dio na području općina Desinić, Veliko trgovišće i Dubravice. Natječaj - Tipski dječji vrtić drvene	**konstrukcije/Ncfsg**	Komunalna naknada Djelatno vrijeme: 08:00 15:00 Pauza: 11:00 11:30 Dobro došli Zavod za urbanizam i izgradnju Osijek
15.	vrijeme: 08:00 15:00 Pauza: 11:00 11:30 Dobro došli Zavod za urbanizam i izgradnju Osijek U okviru mnogobrojnih	**poslova/Ncmpg**	(urbanizam, projektiranje, nadzor, geodezija, gradnja za tržište), posebno izdvajamo obavljanje stručnih
16.	(urbanizam, projektiranje, nadzor, geodezija, gradnja za tržište), posebno izdvajamo obavljanje stručnih	**poslova/Ncmpg**	za Grad Osijek, prigradske općine, Osječko-baranjsku županiju, ministarstva Republike Hrvatske, te vođenja
17.	, nadzor, geodezija, gradnja za tržište), posebno izdvajamo obavljanje stručnih poslova za Grad Osijek, prigradske	**općine/Ncfsg**	, Osječko-baranjsku županiju, ministarstva Republike Hrvatske, te vođenja projekata i izgradnja stambenih,
18.	), posebno izdvajamo obavljanje stručnih poslova za Grad Osijek, prigradske općine, Osječko-baranjsku županiju,	**ministarstva/Ncnsg**	Republike Hrvatske, te vođenja projekata i izgradnja stambenih, poslovnih građevina za tržište. Naše djelatnosti
19.	obavljanje stručnih poslova za Grad Osijek, prigradske općine, Osječko-baranjsku županiju, ministarstva	**Republike/Ncfsg**	Hrvatske, te vođenja projekata i izgradnja stambenih, poslovnih građevina za tržište. Naše djelatnosti Prostorno
20.	stručnih poslova za Grad Osijek, prigradske općine, Osječko-baranjsku županiju, ministarstva Republike	**Hrvatske/Npfsg**	, te vođenja projekata i izgradnja stambenih, poslovnih građevina za tržište. Naše djelatnosti Prostorno

Tablica 17 pokazuje da je većina ključnih riječi (engl. *key word*, stupac 3) dobro označena (osim primjera 17 i 18), te da se uglavnom radi o općim imenicama u genitivu jednine ili množine. No, također je razvidno da primjeri 17-20 predstavljaju isti jezični sadržaj, odnosno jednu rečenicu. To je stoga što se konkordancije prikazuju onim redoslijedom kojim se pojavljuju u korpusu. Želimo li nešto drugačiji redoslijed, potrebno je odabrati opciju *Get a random sample*, koja će prikazati nasumične rečenice iz korpusa.

Možemo, primjerice, pretraživati opće imenice iza kojih slijedi druga imenica u genitivu množine (npr. *Vijeće studenata*), što je prikazano u tablici 18.

Tablica 18: *Konkordancijski niz pretrage* [tag="Nc.*"][tag="N.*cg.*"] *u* hrWaC-*u*

1.	obavlja se na području Grada Klanjca, te općina Tuhelj, Kumrovec, Zagorska sela, Kraljevec na Sutli, te manji dio na	**području/Ncnsl općina/ Ncfpg**	Desinić, Veliko trgovišće i Dubravice. Natječaj - Tipski dječji vrtić drvene konstrukcije Komunalna naknada
2.	poslova za Grad Osijek, prigradske općine, Osječko-baranjsku županiju, ministarstva Republike Hrvatske, te	**vođenja/Ncnsg projekata/ Ncmpg**	i izgradnja stambenih, poslovnih građevina za tržište. Naše djelatnosti Prostorno planiranje Projektiranje
3.	građevina za tržište. Naše djelatnosti Prostorno planiranje Projektiranje Geodezija Stručni nadzor Tehničko	**savjetovanje/Ncnsn investitora/Ncmpg**	Izgradnja stanova, poslovnih prostora i garaža Pored navedenih poslova Zavod za urbanizam i izgradnju d. d. Osijek
4.	Naše djelatnosti Prostorno planiranje Projektiranje Geodezija Stručni nadzor Tehničko savjetovanje investitora	**Izgradnja/Ncfsn stanova/ Ncmpg**	, poslovnih prostora i garaža Pored navedenih poslova Zavod za urbanizam i izgradnju d. d. Osijek obavlja i stručne
5.	da će uskoro biti nadopunjeno. Pozivamo Vas da se registrirate, kako biste dobili pristup skriptarnici te većoj	**količini/Ncfsd sadržaja/ Ncmpg**	te kako biste, ukoliko želite, prvi dobivali sve informacije vezane uz aktualne natječaje i obavijesti koje provodi
6.	evaluaciju je OBAVEZAN za sve redovne studente Zdravstvenog veleučilišta. ‘ ‘. objavljujemo natječaj za	**pokrivanje/Ncnsa troškova/Ncmpg**	sudjelovanja na 1. Konferenciji radnih terapeuta s međunarodnim sudjelovanjem u Zagrebu na temu “ Svakodnevna
7.	, zajedno s drugim glazbeno-scenskim i likovnim umjetnicima. Svrha Festivala je predstavljanje stvaralačkih	**mogućnosti/Ncfpg izvođača/Ncmpg**	programa, šireći poruku da i osobe s invaliditetom trebaju uživati ista prava i obveze poput drugih građana.
8.	trebaju uživati ista prava i obveze poput drugih građana. Studentski zbor Zdravstvenog veleučilišta u	**pratnji/Ncfsl nastavnica/ Ncfpg**	na svečanom otvaranju Foruma struke, koji se održao 17. svibnja 2013. u SD Stjepan Radić, dvorana SKUC Pauk s početkom u
9.	koji se održao 17. svibnja 2013. u SD Stjepan Radić, dvorana SKUC Pauk s početkom u 12 sati. Forum struke organiziralo je	**Vijeće/Ncnsn studenata/ Ncmpg**	Veleučilišta i Visokih škola RH, manifestacija se održavala od 17. do 19. svibnja 2013. godine, a cilj Foruma struke

10.	i klimatizacije Turistička zajednica grada Vrbovca raspisala je natječaj za Zeleni cvijet 2013. godine za sve	**stanovnike/Ncmpa općina/Ncfpg**	Dubrava, Gradec, Farkaševac, Preseka, Rakovec i grada Vrbovca. Tekst natječaja možete preuzeti s naših stranica.
11.	natječaja možete preuzeti s naših stranica. Gradsko izborno povjerenstvo Grada Vrbovca utvrdilo je i objavljuje	**rezultate/Ncmpa izbora/ Ncmpg**	za gradonačelnika Grada Vrbovca, Odluku o održavanju drugog kruga glasovanja u izboru za gradonačelnika Grada
12	Grada Vrbovca, Odluku o održavanju drugog kruga glasovanja u izboru za gradonačelnika Grada Vrbovca te	**Rezultate/Ncmpa izbora/ Ncmpg**	za članice - članove Gradskog vijeća Grada Vrbovca. U četvrtak, 9. svibnja 2013., u Vrbovcu je obilježen Dan Europe. U
13	Petar Zrinski, male gimnastičarke iz II. Osnovne škole Vrbovec te djeca vrbovečkih dječjih vrtića. Predstavnici	**Udruge/Ncfsg branitelja/ Ncmpg**	, policije i vojne policije iz Domovinskog rata predvođeni Antom Jurićem održali su radni sastanak s predsjednikom
14.	s predsjednikom Republike Hrvatske dr. sc. Ivom Josipovićem. Razgovaralo se o brojnim temama za poboljšanje	**statusa/Ncmsg branitelja/Ncmpg**	u društvu. Uz predsjednika Josipovića na sastanku su bili nazočni i savjetnik predsjednika za nacionalnu sigurnost
15.	cijenama. Gradsko izborno povjerenstvo Grada Vrbovca objavljuje Zbirnu listu i pravovaljane kandidature za	**izbor/Ncmsan gradonačelnika/Ncmpg**	Grada Vrbovca, Zbirnu listu i pravovaljane kandidature za izbor članica - članova Gradskog vijeća grada Vrbovca te
16.	listu i pravovaljane kandidature za izbor gradonačelnika Grada Vrbovca, Zbirnu listu i pravovaljane kandidature za	**izbor/Ncmsan članica/ Ncfpg**	- članova Gradskog vijeća grada Vrbovca te Rješenje o određivanju biračkih mjesta na području Grada Vrbovca. Dobro
17.	se stvaraju ustrajnim promicanjem odnosa korektnosti. Poslovne vrijednosti koje promičemo kroz izvrsnost usluge i	**kvalitetu/Ncfsa roba/ Ncfpg**	koje nudimo na tržištu garancija su prepoznatljive kulture našeg poslovanja. Elektromaterijal pokreće svijet Iz
18.	dostignućima praktičnosti korištenja električne energije. Toplinu Vašeg doma u tehničkom smislu stvara	**energija/Ncfsn ljudi/ Ncmpg**	koji u njemu žive. Tvrtka Wellmax d. o. o. je regionalni lider u isporuci elektromaterijala. Dosadašnjim
19.	. Dosadašnjim dvadesetogodišnjim radom etablirala se kao vodeći distributer renomiranih svjetskih proizvođača	**elektro/Ncmsn roba/ Ncfpg**	kao što su FKN, Eurocable, Elka, Elektro-kontakt, Siemens, Hager i brojnih drugih koje možete upoznati na našim
20.	tome pridaju važnost koja je sukladna uzrastu i zrelosti djeteta. Članak 13. Dijete ima pravo na traženje, primanje i	**dobivanje/Ncnsa informacija/Ncfpg**	, usmenih, pisanih, tiskanih i drugih oblika prema svome odabiru. Članak 14. Dijete ima pravo na slobodu izražavanja

Primjer 15 u tablici 18 trebao bi biti označen kao genitive jednine, jer se radi o sinkretizmu oblika, a semantički gledano samo je jedan gradonačelnik. U primjeru 19 imamo imensku složenicu elektro-roba, a ne dvije imenice od kojih je jedna glava, a druga postmodifikator.

Sljedeći primjer pokazuje imenske skupine koje se sastoje od imenice iza koje slijedi prijedlog (npr. *osoba s invaliditetom*), a rezultati su dobiveni pretragom: [tag="N.*"] [tag="S.*"]

Tablica 19: *Konkordancijski niz pretrage* [tag="N.*"] [tag="S.*"] *u* hrWaC-*u*

1.	Uštedite korištenjem plina... Prirodni plin upotrebljava se u	**kućanstvu/Ncnsl za/Sa**	grijanje prostora, pripremu tople vode, termičku obradu hrane, a koristi se i sve više za hlađenje prostora. Prirodni
2.	tople vode, termičku obradu hrane, a koristi se i sve više za hlađenje prostora. Prirodni plin nalazi svoju široku	**potrošnju/Ncfsa u/Sl**	industriji za tehnološke procese, kao gorivo ili kao sirovina, te za grijanje i hlađenje prostora. Kontaktirajte nas
3.	, termičku obradu hrane, a koristi se i sve više za hlađenje prostora. Prirodni plin nalazi svoju široku potrošnju u	**industriji/Ncfsl za/Sa**	tehnološke procese, kao gorivo ili kao sirovina, te za grijanje i hlađenje prostora. Kontaktirajte nas... ZELENJAK
4.	, te za grijanje i hlađenje prostora. Kontaktirajte nas... ZELENJAK PLIN d. o. o. ovlašten je od HRVATSKE STRUČNE	**UDRUGE/Ncfsg ZA/Sa**	PLIN za ispitivanje plinskih instalacija. Distribucija prirodnog plina obavlja se na području Grada Klanjca, te
5.	za grijanje i hlađenje prostora. Kontaktirajte nas... ZELENJAK PLIN d. o. o. ovlašten je od HRVATSKE STRUČNE UDRUGE ZA	**PLIN/Ncmsan za/Sa**	ispitivanje plinskih instalacija. Distribucija prirodnog plina obavlja se na području Grada Klanjca, te općina
6.	. Distribucija prirodnog plina obavlja se na području Grada Klanjca, te općina Tuhelj, Kumrovec, Zagorska sela,	**Kraljevec/Npmsn na/Sl**	Sutli, te manji dio na području općina Desinić, Veliko trgovišće i Dubravice. Natječaj - Tipski dječji vrtić drvene
7.	plina obavlja se na području Grada Klanjca, te općina Tuhelj, Kumrovec, Zagorska sela, Kraljevec na Sutli, te manji	**dio/Ncmsan na/Sl**	području općina Desinić, Veliko trgovišće i Dubravice. Natječaj - Tipski dječji vrtić drvene konstrukcije
8.	dječji vrtić drvene konstrukcije Komunalna naknada Djelatno vrijeme: 08:00 15:00 Pauza: 11:00 11:30 Dobro došli	**Zavod/Ncmsn za/Sa**	urbanizam i izgradnju Osijek U okviru mnogobrojnih poslova (urbanizam, projektiranje, nadzor, geodezija, gradnja
9.	Komunalna naknada Djelatno vrijeme: 08:00 15:00 Pauza: 11:00 11:30 Dobro došli Zavod za urbanizam i izgradnju	**Osijek/Npmsn U/Sl**	okviru mnogobrojnih poslova (urbanizam, projektiranje, nadzor, geodezija, gradnja za tržište), posebno

10.	Zavod za urbanizam i izgradnju Osijek U okviru mnogobrojnih poslova (urbanizam, projektiranje, nadzor, geodezija,	**gradnja/Ncfsn za/Sa**	tržište), posebno izdvajamo obavljanje stručnih poslova za Grad Osijek, prigradske općine, Osječko-baranjsku
11.	(urbanizam, projektiranje, nadzor, geodezija, gradnja za tržište), posebno izdvajamo obavljanje stručnih	**poslova/Ncmpg za/Sa**	Grad Osijek, prigradske općine, Osječko-baranjsku županiju, ministarstva Republike Hrvatske, te vođenja
12.	županiju, ministarstva Republike Hrvatske, te vođenja projekata i izgradnja stambenih, poslovnih	**građevina/Ncfpg za/Sa**	tržište. Naše djelatnosti Prostorno planiranje Projektiranje Geodezija Stručni nadzor Tehničko savjetovanje
13.	Geodezija Stručni nadzor Tehničko savjetovanje investitora Izgradnja stanova, poslovnih prostora i	**garaža/Ncfpg Pored/Sg**	navedenih poslova Zavod za urbanizam i izgradnju d. d. Osijek obavlja i stručne poslove u svezi pripreme građevinskog
14.	nadzor Tehničko savjetovanje investitora Izgradnja stanova, poslovnih prostora i garaža Pored navedenih poslova	**Zavod/Ncmsn za/Sa**	urbanizam i izgradnju d. d. Osijek obavlja i stručne poslove u svezi pripreme građevinskog zemljišta, komunalnog
15.	, poslovnih prostora i garaža Pored navedenih poslova Zavod za urbanizam i izgradnju d. d. Osijek obavlja i stručne	**poslove/Ncmpa u/Sl**	svezi pripreme građevinskog zemljišta, komunalnog opremanja, prometne infrastrukture i komunalne naknade za Grad
16.	poslove u svezi pripreme građevinskog zemljišta, komunalnog opremanja, prometne infrastrukture i komunalne	**naknade/Ncfsg za/Sa**	Grad Osijek po godišnjem ugovoru. dobrodošli na Internet stranice studentskog zbora Zdravstvenog veleučilišta u
17.	svezi pripreme građevinskog zemljišta, komunalnog opremanja, prometne infrastrukture i komunalne naknade za Grad	**Osijek/Ncmsan po/Sl**	godišnjem ugovoru. dobrodošli na Internet stranice studentskog zbora Zdravstvenog veleučilišta u Zagrebu. Od sada
18.	naknade za Grad Osijek po godišnjem ugovoru. dobrodošli na Internet stranice studentskog zbora Zdravstvenog	**veleučilišta/Ncnsg u/Sl**	Zagrebu. Od sada, ovdje ćete moći pronaći sve informacije vezane uz djelovanje i aktivnosti zbora te važne obavijesti
19.	skripti, i biti što aktivniji na forumu te Vam odgovarati na sve eventualne upite i poteškoće. Ovim putem Vas molimo za	**razumijevanje/Ncnsa zbog/Sg**	trenutnog manjka sadržaja, što se nadamo da će uskoro biti nadopunjeno. Pozivamo Vas da se registrirate, kako biste
20.	sve informacije vezane uz aktualne natječaje i obavijesti koje provodi Studentski zbor. Podsjećamo Vas i pozivamo na	**evaluaciju/Ncfsa putem/Sg**	studomata. Elektronička evaluacija jamči nam brzu, jednostavnu i potpuno anonimnu ocjenu nastave i nastavnika, te

Ovakve sintaktičke pretrage složene su i predstavljaju izazov računalnojezikoslovnim alatima jer je teško razgraničiti kada prijedložna skupina koja slijedi postmodificira imenicu (npr. *osoba s invaliditetom*), a kada ima ulogu priloga u rečenici, odnosno kada je riječ o sintaksi skupine, a kada o sintaksi (su)rečenice.

Iz primjera u tablici 20 vidljivo je da samo primjeri *industrija za tehnološke procese*, *udruga za plin*, *Kraljevec na Sutli*, *Zavod za urbanizam i izgradnju*, *poslove u svezi pripreme građevinskog zemljišta*, *naknade za grad Osijek* odgovaraju definiciji postmodifikacije. Konkordancije iz tablice 20, dakle, mogu biti polazište za sužavanje pretrage na način da definiramo da iza jedne imenice slijedi prijedlog, a potom pridjev ili druga imenica.

Tablica 20: *Konkordancijski niz pretrage* [tag="N.*"] [tag="S.*"] [tag="A.*"]? [tag="N.*"] *u* hrWaC-*u, slučajni uzorak*

1.	No, u tom naumu ga je spriječio fra Stanko Litre koji ga je udaljio iz crkve. Međutim, momak se ponovno vratio u	**crkvu/Ncfsa za/Sa vrijeme/Ncnsa**	mise i to dva puta pa smo ga opet izbacili, a nakon toga je pozvana i policija koja ga je privela u prostoriju PP Makarska
2.	sposobnosti i na stvaranje pozitivne slike o sebi, stvaranje prijateljstava, samostalnosti, tjelesne aktivnosti,	**ljubavi/Ncfsg prema/Sl sportu/ Ncmsl**	sportskom ponašanju i zdravom životu. Osnaženje tjelesnih sposobnosti i uspjeh u ovladavanju tijelom povezani su s
3.	razvoja maslinarstva i poljoprivrede na ovom području vidi u povezanosti sa agroturizmom: Facebook Kontakti	**Suradnja/Ncfsn s/ Si Ministarstvom/ Ncnsi**	unutarnjih poslova (MUP) i razvoj obrazovnog modela za policijske službenike. Istraživanje Mladi i opijanje Grad
4.	tvrtka koja se bavi ugradnjom i modernizacijom sustava grijanja te ugrađuje solarne panele i toplinske crpke na	**zgrade/Ncfpa u/Sl potrazi/Ncfsl**	je za novim tehnologijama koje će pridonijeti razvoju budućih sustava. Tvrtka traži tehnike koje se mogu primijeniti
5.	foruma (WEF) o informacijskoj tehnologiji, dok je u godini prije bila na 45. mjestu, objavilo je danas Nacionalno	**vijeće/Ncnsn za/ Sa konkurentnost/ Ncfsa**	. Trenutna ekonomska situacija i zbivanja na hrvatskom tržištu tvrtkama su nametnuli zahtjev za smanjivanjem
6.	rada na računalu i intervju održat će se dana 29.04.2013. godine s početkom u 08:30 sati u općinskoj vijećnici Općine	**Civljane/Npfsg na/ Sl adresi/Ncfsl**	Kod doma 3, 22310 Kijevo. Kandidatkinja je dužna sa sobom imati važeću osobnu iskaznicu. OBAVIJEST Obavještavaju se
7.	na papiru. Dobit ćete: kostur čestitke u obliku jaja, naslovnicu na kojoj piše: Sretan Uskrs, zeku, košaru i 5 malenih	**jaja/Ncnpg u/Sl različitim/Agpfply bojama/Ncfpl**	. Mali trik: ja uvijek rubove elementima koje ručno režem pomalo izližem. Em to samim elementima da duše jer se tako i
8.	prava radnika predstavljala kao pitanje života i smrti stižu obavijesti o povećanju primanja određenih kategorija	**radnika/Ncmpg s/ Si objašnjenjem/ Ncnsi**	da radnici obuhvaćeni ovom povišicom imaju doista puno posla. Pozdravljamo svaku namjeru adekvatnog vrednovanja

9.	da stoji na mjestu. (puder ima zaštitni faktor 15). Mislim da se na slici lijepo vidi da paleta djeluje:) Meni je osobno	**korektor/Ncmsn u/ Sl boji/Ncfsl**	kože prekremast pa mi se uvuče u bore oko očiju (a vidite da ih ima) pa umjesto njega koristim od Bourjois Heatlhy mix
10.	organizacije, poslovnih aktivnosti i portfelja. Najvećim dijelom neznalice? Tri faze restrukturiranja Iskreni	**odnos/Ncmsn s/Si ljudima/Ncmpi**	Vrijeme je za promjenu smjera - Sa zavoda za zapošljavanje u proizvodnju, a ne obrnuto Kako se približavaju izbori,
11.	toj najsiromašnijoj karipskoj zemlji pogođenoj razornim potresom. DiCaprijeva donacija uplaćena je u Zakladu	**Clinton-Bush/ Npmsn za/Sa Haiti/Npmsan**	, izvijestile su američke TV mreže. Američki predsjednik Barack Obama imenovao je bivše predsjednike Billa Clintona
12.	Pavelicem??? Sto trazi taj 8. Tip, Spasioc Svijeta kod Pavelica??? 9. Klerikalci, ta zadnja drustvena Bagra suti o	**Genocidima/ Ncmpl u/Sl Bibliji/ Ncfsl**	David je jedan od najvecih ratnih Zlocinaca u Starom Testamentu 1. Biblija vrvi od njegovih Zlocina, Ratova, i ratnih
13.	Darkwooda. Bila je to 18. po redu završnica Kupa RH koja je odigrana u nedjelju 25. rujna u Upravnoj zgradi Marine	**Kaštela/Npmsg u/ Sl Kaštel/Npmsl**	Gomilici, a nastupilo je 17 ekipa u muškoj i 8 u ženskoj konkurenciji. ČAK 183 IGRAČA Odličan dan V. Lešića, povratak
14.	nalaze se u gradu i okolici. Oni su ispisali bogatu povijest ovoga grada. Iskopavanja iz 1997. godine potvrđuju	**život/Ncmsan na/ Sl širem/Agpnsly području/Ncnsl**	grada već u neolitiku, oko 5500 godina prije Krista, koji kroz dugu i burnu povijest traje do današnjih dana. U rimsko
15.	. Triba znati da je to jedna visoko centralizirana strančica s jako malo autonomije na nižim razinama, tj. Lesar bi ih	**poviša/Ncmsg za/ Sa jaja/Ncnpa**	kad bi koalirali s HDZ-om. Boj se konja i kad darove nose Ovo je jedina stranka koja nikad nije iskazala povjerenje
16.	Atletico Madrid i Valencija. Atletičari čvrsto drže treću poziciju, a Valencija se nada četvrtom mjestu i još jednoj	**sezoni/Ncfsl u/Sa LP/Npmsan**	. Šišmiši se već dugo šuljaju prema četvrtoj poziciji, a trenutno im fale dva boda da dostignu Sociedad. Protiv
17.	raditi duže od šest mjeseci. To je i razlog zbog kojeg je Vlada DZS-u odobrila zapošljavanje do 400 radnika na rok do šest	**mjeseci/Ncmpg uz/Sa mogućnost/ Ncfsa**	dodatnog angažiranja do 200 osoba na još četiri mjeseca. Kao što smo već pisali, DZS će probiti zakonske rokove i popis
18.	vrlo kratko. Stručna izobrazba PUZS RIJEKA Potaknuti aktualnom situacijom u DINA Petrokemiji d. d., vezano uz moguće	**nesreće/Ncfpa sa/Si opasnim/ Agpfpiy tvarima/ Ncfpi**	, te posljedicama koje mogu prouzročiti, Krizni stožer ministarstva zdravlja predložio je Stožeru zaštite i
19.	singla “ Hipnotiziran “, obitelj i prijatelji prerano preminulog kralja hrvatskog funka Dine Dvornika nastavili su	**poslove/Ncmpa oko/Sg izdavanja/ Ncnsg**	albuma Pandorina kutija koji bi se na tržištu trebao naći 30. studenoga. Na radijskim postajama uskoro će se početi
20.	Narode hrvatski, bre, meni vas je žao. Vi ste uskoro samo jadna fleka na polju interesa koje mudri stranci razvijaju	**ovde/Ncmpa uz/Sa pomoć/Ncfsa**	vaših kretena, izdajnika i blesavih kurvi, posebno medijskih. Ili ćete se sami izdefinirati, i pronaći tome srodnu i

Prilagođena pretraga polučila je nepotpune rezultate jer samo u deset od navedenih dvadeset primjera priložna skupina koja slijedi postmodificira imenicu: *ljubavi prema sportu*, *suradnja s ministarstvom*, *Vijeće za konkurentnost*, *jaja u različitim bojama*, *korektor u boji*, *odnos s ljudima*, *Genocidima u Bibliji*, *život na širem području*, *nesreće sa opasnim tvarima*, *poslove oko izdavanja.*

Ovi problemi, kao što je, primjerice, tzv. *prepositional phrase attachment ambiguity* (Delecraz i dr., 2021), ne mogu se riješiti nikako osim ručnom pretragom, a čak ni dubinski parseri nisu uspješni u razrješavanju ovoga problema (tablica 21). Dodatak, dopuna, prilog i sl. često su pogrešno označeni u korpusu, a pogotovo nisu razjašnjeni u hrvatskome jeziku, što je problematično jer parseri polaze od neke sintaktičke teorije, a ako iste nema ili ako nema konsenzusa oko toga što se smatra dopunom ili dodatkom, teško je provesti korpusno utemeljena istraživanja sintaktičkih struktura. No, projektom *Sintaktička i semantička analiza dopuna i dodataka u hrvatskom jeziku* (*SARGADA*) voditeljice M. Birtić definirat će se „jasni i precizni kriteriji za razlikovanje dopuna i dodataka u hrvatskom jeziku te ih primijeniti u izgradnji sintaktičkog repozitorija hrvatskoga jezika koji bi bio vrijedan resurs za unaprjeđenje alata za strojnu obradu jezika te za proučavanje i podučavanje hrvatskog jezika”[111].

Tablica 21: *Dubinsko parsanje imenske skupine* život na širem području grada

Surface	Tags	Lemma	Dep parse - gov / func	Paragraph	Sentence	Token	Start char	End char	
1.	život	Ncmsn	život	0 / root	1	1	1	1	5
2.	na	Sl	na	4 / case	1	1	2	7	8
3.	širem	Agpnsly	širok	4 / amod	1	1	3	10	14
4.	području	Ncnsl	područje	1 / nmod	1	1	4	16	23
5.	grada	Ncmsg	grad	4 / nmod	1	1	5	25	29

Želimo li, primjerice, iz korpusa izlistati imenske skupine u kojima odnosna surečenica postmodificira glavu skupine (imenicu), jezikoslovac treba osmisliti pravila, kao npr. [tag="Nc.*"] [lemma="koji|što|kakav|tko|čiji|kolik|komu|čemu(…)"]. Pravilo koje smo postavili dat će rezultate prikazane u tablici 22. Na temelju tih rezultata pretragu možemo dalje sužavati.

[111] Izvor: *Sintaktička i semantička analiza dopuna i dodataka u hrvatskom jeziku (SARGADA) - Institut za hrvatski jezik i jezikoslovlje.*

Tablica 22: *Konkordancijski niz pretrage* [tag="Nc.*"] [lemma="koji|što|kakav|tko|čiji|kolik|komu|čemu"] *u* hrWaC-*u*

1.	većoj količini sadržaja te kako biste, ukoliko želite, prvi dobivali sve informacije vezane uz aktualne natječaje i	**obavijesti/Ncfpn koje/Pi-fpa**	provodi Studentski zbor. Podsjećamo Vas i pozivamo na evaluaciju putem studomata. Elektronička evaluacija jamči
2.	godine 2012. / 2013. Evaluacija nastave za studente redovitih studija održati će se u dvorani 304, Ksaver 209 prema	**rasporedu/Ncmsl čiji/Pi-msn**	link Vam slijedi u nastavku. Dolazak na elektroničku evaluaciju je OBAVEZAN za sve redovne studente Zdravstvenog
3.	iskustava i znanja u radu s osobama s invaliditetom. F = M je tradicionalna međunarodna manifestacija urbane	**kulture/Ncfsg čiji/Pi-mpn**	program izvode osobe s invaliditetom, zajedno s drugim glazbeno-scenskim i likovnim umjetnicima. Svrha Festivala
4.	, Domaća radinost Ljiljane Stanko i OPG Lisjak. Pozivamo građane da dođu i pogledaju ponudu kvalitetnih hrvatskih	**proizvoda/Ncmpg koji/Pi-mpn**	će se tom prilikom moći kupiti po prigodnim cijenama. Gradsko izborno povjerenstvo Grada Vrbovca objavljuje Zbirnu
5.	naše tvrtke i asortiman ponude u interesu stvaranja partnerskih odnosa s Vama. Tako možete upoznati dio naše poslovne	**tradicije/Ncfsg koja/Pi-fsn**	govori o trajnim vrijednostima koje se stvaraju ustrajnim promicanjem odnosa korektnosti. Poslovne vrijednosti
6.	interesu stvaranja partnerskih odnosa s Vama. Tako možete upoznati dio naše poslovne tradicije koja govori o trajnim	**vrijednostima/Ncfpl koje/ Pi-fpn**	se stvaraju ustrajnim promicanjem odnosa korektnosti. Poslovne vrijednosti koje promičemo kroz izvrsnost usluge i
7.	tradicije koja govori o trajnim vrijednostima koje se stvaraju ustrajnim promicanjem odnosa korektnosti. Poslovne	**vrijednosti/Ncfpa koje/Pi-fpa**	promičemo kroz izvrsnost usluge i kvalitetu roba koje nudimo na tržištu garancija su prepoznatljive kulture našeg
8.	ustrajnim promicanjem odnosa korektnosti. Poslovne vrijednosti koje promičemo kroz izvrsnost usluge i kvalitetu	**roba/Ncfpg koje/Pi-fpa**	nudimo na tržištu garancija su prepoznatljive kulture našeg poslovanja. Elektromaterijal pokreće svijet Iz
9.	dostignućima praktičnosti korištenja električne energije. Toplinu Vašeg doma u tehničkom smislu stvara energija	**ljudi/Ncmpg koji/Pi-mpn**	u njemu žive. Tvrtka Wellmax d. o. o. je regionalni lider u isporuci elektromaterijala. Dosadašnjim
10.	drugoga. Članak 12. Dijete ima pravo na slobodno izražavanje svojeg mišljenja i iskustva te da odrasli tome pridaju	**važnost/Ncfsn koja/Pi-fsn**	je sukladna uzrastu i zrelosti djeteta. Članak 13. Dijete ima pravo na traženje, primanje i dobivanje informacija,

11.	Dijete ima pravo na zaštitu od zlostavljanja, zanemarivanja i svih oblika tjelesnoga i duševnoga nasilja. Članak 20.	**Dijete/Ncnsn koje/Pi-nsn**	je trajno ili privremeno bez obitelji ima pravo na zamjensku pomoć: smještanjem kod hranitelja, usvojenjem, ili
12.	posebna zaštita, skrb i humanitarna pomoć. Članak 23. Dijete s duševnim ili tjelesnim nedostatkom ima pravo živjeti u	**uvjetima/Ncmpl koji/Pi-mpn**	mu osiguravaju dostojanstvo, potiču vlastite mogućnosti i olakšavaju djelatno sudjelovanje u zajednici. Članak
13.	zaštitu, na liječenje i oporavak od bolesti, na primjerenu hranu, pitku vodu i na čistoću okoliša. Članak 25.	**Dijete/Ncnsn koje/Pi-nsn**	je odlukom nadležne vlasti stavljeno pod nadzor u odgovarajuću ustanovu ima pravo tražiti da stručnjaci povremeno
14.	povremeno provjere i, ako je potrebno, utječu na smještaj i druge okolnosti u kojima dijete živi. Članak 26.	**Dijete/Ncnsn kojemu/Pi-msd**	je to potrebno ima pravo na socijalnu skrb, na povlastice koje će mu omogućiti socijalnu sigurnost i socijalno
15	smještaj i druge okolnosti u kojima dijete živi. Članak 26. Dijete kojemu je to potrebno ima pravo na socijalnu skrb, na	**povlastice/Ncfpa koje/Pi-fpa**	će mu omogućiti socijalnu sigurnost i socijalno osiguranje. Članak 27. Dijete ima pravo na životni standard
16.	za odgovoran život u slobodnoj zajednici u duhu razumijevanja, mira i prijateljstva među svim narodima. Članak 30.	**Dijete/Ncnsn koje/Pi-nsn**	pripada etničkoj manjini ima pravo na upoznavanje vrednote svoje kulture, ispovijedanje svoje vjere i na uporabu
17.	u željenim stvaralačkim, kulturnim, sportskim i zabavnim aktivnostima. Članak 32. Dijete ima pravo na zaštitu od	**rada/Ncmsg koji/Pi-msn**	je pretežak za dijete od posla koji je opasan, ometa naobrazbu, šteti zdravlju i koji ekonomski izrabljuje dijete.
18.	, sportskim i zabavnim aktivnostima. Članak 32. Dijete ima pravo na zaštitu od rada koji je pretežak za dijete od	**posla/Ncmsg koji/Pi-msn**	je opasan, ometa naobrazbu, šteti zdravlju i koji ekonomski izrabljuje dijete. Članak 33. Dijete ima pravo na zaštitu
19.	od sudjelovanja u oružanim sukobima i od novačenja u oružane postrojbe dok ne navrši najmanje 15 godina. Članak 39.	**Dijete/Ncnsn koje/Pi-nsn**	je žrtva bilo kakva nehaja, izrabljivanja, zlouporabe ili oružanog sukoba ima pravo na odgovarajući tjelesni i
20.	ili oružanog sukoba ima pravo na odgovarajući tjelesni i duševni oporavak i ponovno uklapanje u društvo. Članak 40.	**Dijete/Ncnsn koje/Pi-nsn**	je optuženo da je prekršilo zakon ima pravo smatrati se nedužnim sve dok mu se krivnja ne dokaže u skladu sa zakonom, a

U hrvatskome postmodifikatori glave imenske skupine mogu biti prijedložna skupina, zatim druga imenska skupina (i genitivna skupina), finitna surečenica, pridjevska skupina i priložna skupina, dok je najrjeđi oblik postmodifikacije nefinitna surečenica. No, želimo li saznati koji su najčešći postmodifikatori, korpusni podaci koje dobijemo mogu biti samo okvirni pokazatelj (v. tablicu 23).

Tablica 23: *Postmodifikatori glave imenske skupine; rezultati dobiveni pretraživanjem* HNK-*a*

Oblik postmodifikacije	**CQL**	**RF hrWaC**	**RF Riznica**	**Primjer**
imenica (= glava) + (genitivna) imenska skupina	[tag="N.*"][tag="A.*|M.*"][tag="N.*g*"]	12367,09	15118,17	*istraživanja različitih sastavnica; izdanje autorove studije; kolekciju internih skripti*
imenica (= glava) + prijedložna skupina	[tag="N.*"][tag="S.*"] [tag="A.*|M.*"]? [tag="N.*"]	26753,82	29323,47	*potrošnju u industriji, zavod za urbanizam; pravo na uvođenje moderniteta*
imenica (=glava) + odnosna surečenica bez prijedloga	[tag="N.*"] [tag="S.*"] [lemma="koji|što|kakav|tko|čiji|komu|kojemu"]	5714,12	7038,14	*režimu koji je, slobode kakve je, miljea kojim je*
imenica (=glava) + odnosna surečenica uvedena prijedlogom	[tag="N.*"] [tag="S.*"] [lemma="koji|što|kakav|tko|čiji|komu|kojemu"]	1382,67	1455,28	*okolnosti u kojima se javljaju…, atmosferu u kojoj se odvijaju…, postavke na kojima je utemeljena… zakonima u kojima su propisana*

Za postmodifikaciju nefinitnom surečenicom nismo dobili konkretne podatke iz korpusa. Naime, zbog prirode hrvatskoga jezika, vrlo je teško dobiti

jednoznačne podatke iz korpusa i postaviti pretragu tako da obuhvaća samo jednu imensku skupinu.

U ovome dijelu pokazali smo da korpusi mogu biti važan izvor hrvatskim gramatikama, po uzoru na Biber i dr. (1999) i Carter i McCarthy (2006) jer hrvatski nema korpusom vođenu ili korpusno utemeljenu gramatičku tradiciju. Primjerice, hrvatske gramatike nefitine surečenice ne smatraju nezavisnima, a to je tema koju treba preispitati.

7.2.2. VEZNICI

Osim na razini riječi i leksičkih jedinica, skupine i sintaktičkih jedinica, korpusi mogu dati uvid u suprasintaktičke strukture i omogućiti prepoznavanje karakteristika žanra ili diskursa koje nadilaze razinu riječi (npr. veznike i diskursne oznake itd.)

Koordinacijske i subordinacijske veznike dobit ćemo jednostavno pretragom [tag="Cc"] i [tag="Cs"], a ako tražimo veznike samo na početku rečenice, ispred ćemo dodati oznaku <s>.

Tablica 24: *Konkordancijski niz pretrage* <s> [tag="Cc"] *u* hrWaC-*u*

1.	europskih zemalja ukoliko se prije toga ne raspadne.	**No/no/Cc**	jeli sve to dovoljno. Sjetite se veselog zagrljaja
2.	, kao i oni koji su sve to iskusili na vlastitoj koži.	**A/a/Cc**	sve to doista upućuje na jedan jedini odgovor: Ne I to bez
3.	s nekim kalupom izlio je iz betona (ili nečeg drugog ovo:	**I/i/Cc**	to je netko, naravno, stavio oko kuće. No, da taj ne bi bio sam
4.	jedinica u svojim povijesnim granicama.	**No/no/Cc**	, uslijedile su i srpske protumjere. Tako su se uspostavi
5.	sadržaji odlično su oblikovani, bogati i zanimljivi.	**No/no/Cc**	, postavlja se nekoliko pitanja koja traže odgovore i
6.	s majstorom. O veličini pločica već smo dali uputu.	**No/no/Cc**	, to nije tako čvrsto pravilo. Naime, i velike i male pločice
7.	gdje je uz Papin odar prošlo više od 3 milijuna vjernika.	**No/no/Cc**	, Vječni grad je bio i još nešto: srce svijeta. U plaču i tuzi
8.	ledger@net.hr slike ne tražim, niti bih vam ih dao.	**Ali/ali/Cc**	... ni to nije sve, ja molim i: ne trpati želje, čestitike i
9.	, nažalost lopta je za metar dva bila neprecizna.	**A/a/Cc**	onda drama u sudačkoj nadoknadi prvog poluvremena. Pehar
10.	svaki krivi potez može završiti na naslovnicama tabloida.	**I/i/Cc**	dok smo još uvijek privrženi Natalie Portman i Christini

11.	u međuvremenu imao brže stope rasta od prosjeka Hrvatske.	**No/no/Cc**	, ipak nismo blizu nacionalnog prosjeka BDP-a po glavi
12	sastanak s nadležnim tijelima u Ministarstvu turizma.	**A/a/Cc**	kako saznajemo od njega nakon sastanka, natječaj za tu
13.	pripisuju Božjem zahvatu i Božjoj ljubavi. A mi, Hrvati?	**I/i/Cc**	mi imamo čudesnih iskustava i to upravo iz nedavnih ratnih
14.	". i to niti jednog blogera dusebriznika ne zabrinjava. ()	**a/a/Cc**	ta provokacija (blago receno), odnosno klica razdora i zla
15.	naletu uspjeli su srušiti nekih pedesetak kuća.	**A/a/Cc**	onda je akcija zamrla. Obrazloženje: počinje turistička
16.	, Iskušanim i pravim, Obnavlja se naša mladost.	**Ali/ali/Cc**	vaj Kad umre prijatelj star, Nov mora u tvoje srce ući.
17.	odrasla čovjeka (i nekih životinja, npr. jarca i koze).	**I/i/Cc**	lat. barba, brada, znači zapravo dlake na bradi, jer i
18.	ne, jer vas već sam smrad pokvarenosti odvlači od te hrane.	**A/a/Cc**	zamislite nas ljude kada smo pokvareni Ovdje ne mislim na
19.	postupka? I osvete ženi koja se svidjela njenu mužu?	**Ali/ali/Cc**	zašto u ralje javnosti baca i sve troje svoje djece,
20.	idem.. pozdrav RIBARIĆIMA Napisao yinx @ 17.11.2008 11:51	**Pa/pa/Cc**	dobro dragi prijatelji HDZ-ovci. Zar vi nemate nikog

Tablica 25: *Konkordancijski niz pretrage* <s> [tag="Cs"] *u* hrWaC-*u*

1.	internetski promet preko CARNet mreže i najam modema.	**Ako/ako/Cs**	linija ne omogućava maksimalnu brzinu, Iskon će
2.	i nastaviti pružati širok spektar ruralnih usluga.	**Da/da/Cs**	bi se to dogodilo, poljoprivrednici moraju imati jače
3.	li majka uzimati lijekove i kakve da bi djetetu bilo bolje.	**Kako/kako/Cs**	funkcionira metoda? Kod punkcije pupčane vrpce liječnik
4.	, neoštećene plodove, sa kratko podrezanom peteljkom.	**Iako/iako/Cs**	je plod relativno tvrd kod berbe, kora se lako oštećuje pa se
5.	kriterije: da do 1. ožujka 2013. g. nije napunio 40 godina	**da/da/Cs**	je član stručnog udruženja koje je član EFLM da je prijavio
6.	, Obitelj Soprano, Igru prijestolja, Izgubljene i Žicu.	**Ako/ako/Cs**	ste " spoilerofob ", smatrajte ovo upozorenjem.
7.	, Curly nema drugog izbora nego vratiti Kida u zatvor.	**Kako/kako/Cs**	putovanje napreduje, Kidu se počinje sviđati Dallas.
8.	boluju od glaukoma, rizik je pet do sedam puta veći...	**Da/da/Cs**	bismo razumjeli zašto se nešto pokvarilo i kako to
9	stranica, servisa, aplikacija i drugih mrežnih entiteta.	**Kao/kao/Cs**	i većina ključnih web protokola, kao što su SOAP koji je

10.	proglašavaju favoritom izbornog nadmetanja u HDZ-u.	**Ako/ako/Cs**	uspije u svom naumu i postane novi šef HDZ-a, bivši bi se
11.	gledali kontinuirani rast do završnih 7,285 kuna za euro.	**Iako/iako/Cs**	je u središtu zbivanja trebala biti objava rezultata
12.	su uprihodovali cca 120 milijuna kuna, a URIHO tek mrvice.	**Da/da/Cs**	bi uz takvo poslovanje mogli opstati i isplaćivati
13.	bi po mogućnosti biti posve tamna i zvučno izolirana.	**Ukoliko/ ukoliko/Cs**	nemamo mogućnosti da je otpuno zatamnimo, dobro je
14.	da ce do tog i tog datuma u mjesecu isplatit posloprimca.	**Ako/ako/Cs**	ga ne isplati zakonski snosi sankcije, odnosno trebao bi.
15.	.Ali nisam, nisam joj ni prilazio kamoli letio za njom.	**Da/da/Cs**	me u vezi trazila ej, ja hocu ici tu i hocu radit to pa ja kao lud
16.	da je potrebno i da ostvarujem ciljeve koje sam postavila.	**Da/da/Cs**	sam ostala u angažmanu danas bih imala 31 godinu staža i već
17.	i dobio nadimak " Punisher ", prema osvetniku iz stripova.	**Kako/kako/Cs**	kaže Volkov, do sada je imao više od sto prometnih nezgoda.
18.	i do pola litre votke u kratkom vremenskom razdoblju.	**Iako/iako/Cs**	dečki još uvijek mogu popiti količinski više alkohola
19.	znanja i iskustva da unaprijedi poslovanje Crobenza.	**Budući/budući/ Cs**	da su u Agenciji za zaštitu tržišnog natjecanja
20.	novim kapitalistima. - Čekajte, što znači pogodovati?	**Ako/ako/Cs**	imate osnovni dokument, a to je Generalni plan, ako imate i

Kao koordinacijske veznike *SkE* prepoznaje sljedeće veznike: *i*, *da*, *a*, *kao*, *ili*, *te*, *kako*, *ali*, *jer*, *pa*, *što*, *ako*, *nego*, *dok*, *kad*, *no* itd. Iako prema MULTEXT-East v. 4 veznici u hrvatskome imaju dva atributa (koordinacijski/subordinacijski i jednostavni/složeni), u praksi takva podjela ne vrijedi jer, primjerice, pretrage [tag="Ccs"] i [tag="Ccs"], tj. koordinacijski jednostavni i složeni veznici, daju isti rezultat, što znači da korisnik automatski može pretraživati veznike samo kao subordinacijske i koordinacijske, a ne jednostavne i složene, ili ručno može filtrirati potonje, napisati *gramatiku za crpljenje naziva* (v. 3.) ili smisliti pravilo kojim bi računalu dao naredbu da pronađe sljedeće složene veznike: *pošto*, *zato što*, *zbog toga što*, *uslijed toga što*, *zahvaljujući tome što*, *kao što*, *osim što*, *nego što*, *umjesto što*, *nakon što*, *prije nego što*, *samo što*. Stoga valja sastaviti upit kao što je primjerice: [!tag="N.*|V.*|A.*|R.*|M.*"]{1,5}[word="što"], što bi se moglo iščitati kao: „Pronađi sve pojave riječi što

ispred koje slijedi 1 do 5 riječi koje nisu imenica, glagol, pridjev, prilog ili broj.“ Rezultati takve pretrage prikazani su u tablici 26.

Tablica 26: *Konkordancijski niz pretrage*
[!tag="N.*|V.*|A.*|R.*|M.*"]{1,5}[word="što"] *u* hrWaC-*u*

1.	baš previše povjerenja djeluje više kao boja ili lak	**,/,/Z nego/nego/ Cc kao/kao/Cs nešto/nešto/ Pi3n-a što/što/ Pi3n-n**	može izdržati bilo kakav udarac. Naime, takve kacige
2.	. Podupirući članovi plaćaju nešto veću članarinu	**,/,/Z za/za/Sa što/ što/Pi3n-a**	dobivaju povlastice, npr: objav-ljivanje novosti u našem
3.	i onda im je to da ne kažem evo neka predrasuda ili stereotip	**o/o/Sl njima/ oni/Pp3-pl što/ što/Cs**	će ljudi govoriti ‘, rekao je Duško. Nura, Duško i Čazim bili
4.	staža omogućuje stjecanje većeg iskustva na radnom mjestu	**,/,/Z što/što/Cs**	pak rezultira boljim postignućima. “ Primjereno je
5.	palo na tako niske grane. Foto-grafiji se pristupa kao robi	**,/,/Z za/za/Sa što/ što/Pi3n-a**	su dobrim dijelom krivi proizvođači fotografske tehnike i
6.	, vrlo čisti. No, živjeti s ljenjiv-cem? To je pak druga priča	**././Z Osim/osim/ Cs što/što/Cs**	će vam glavobolju donijeti sama legalizacija, jer držite
7.	mostove prijateljstva s vršnjacima iz Slaavonskog Broda	**i/i/Cc što/što/ Pi3n-n**	će svojim člankom i istraživanjem informirati širu
8.	vaš životni vijek na temelju vaših dnevnih odluka, koliko	**i/i/Qo što/što/Cs**	jedete te koliko često vježbate. Dobiveni “ životni broj “
9.	prilagoditi do točnosti od 0,5 piksela okomito i vodoravno	**,/,/Z što/što/ Pi3n-n**	je od ključne važnosti u post-avama s više projektora, u
10.	. Sa strankom na vlasti i na čelu s potpuno promašenim	**,/,/Z što/što/ Pi3n-n**	se tiče iskustva i sposobnosti ali prepotentnim bez
11.	., 21:22 Istina je ono što dolazi iz mozga. Ljubav je	**ono/on/Pp3nsn što/što/Pi3n-n**	dolazi iz žlijezda pod kontrolom mozga. Za Pravu Ljubav se
12.	i naroda iz kojeg dolazi jer tu nastupaju konvencije (	**kao/kao/Cs što/ što/Cs**	bi Hobbes rekao, sa rođenjem smo protiv naše volje
13.	zna što bi sa sobom. Frustracija je, ipak, bila prejaka	**i/i/Cc nakon/ nakon/Cs što/ što/Cs**	je nekoliko stotina puta opsovao danske suce i pomislio

14.	posve nenadano u ponor bez svjetla grabim i tražim bilo što	**za/za/Sa što/što/ Pi3n-a**	se mogu uhvatiti, bilo što što bi me moglo zaustaviti u tom
15.	2 mm debljine kožnog nabora bez ikakvih kliničkih znakova,	**kao/kao/Cs što/ što/Cs**	je npr. difuzni ili jaki edem, eksudacija, nekroza, bol ili
16.	? Vjerojatno ima. Ali za njih sigurno nećeš čuti, jer to nije	**ono/on/Pp3nsn što/što/Pi3n-a**	svijet traži koliko god tu i tamo povremeno netko zakukao
17.	. Iscrpnije... Kineski izvoz u prosincu je ojačao više	**nego/nego/Cc što/što/Pi3n-n**	se očekivalo, pokazali su u četvrtak najnoviji podaci
18.	za dom i kućanstvo na istoj adresi. Piling i maska od jagoda	**Osim/osim/Cs što/što/Cs**	su ukusne i bogate vitaminima A i C te kalcijem i željezom,
19.	i neće se bitno razlikovati od konačnih podataka. IT:	**I/i/Cc što/što/ Pi3n-a**	kažu brojke? Kakva je bila 2008? Žitnik: U 2008. godini
20.	slika o tome kakvi su uvjeti života na tim prostorima	**,/,/Z što/što/ Pi3n-n**	je bitna pretpostavka za ostvarivanje svih pa i manjinskih

Kvantitativna korpusna analiza ovisi o načinu na koji su dobiveni podaci te je ovisna o reprezentativnosti korpusa te pomnom razumijevanju jezika za postavljanje složenih upita (Meurers i Müller, 2009).

Ovdje se posebno osvrćemo na pitanje zašto su korpusi zanimljivi za ispitivanje sintaktičkih pojava, što su jasno saželi Meuers i Müller (2009). Da bismo proučavali sintaktičku pojavu, primjere moramo svesti na svojstva koja su relevantna za jezična pitanja koja istražujemo, te ih je potrebno varirati da bismo istražili gramatičke odnose ili relacije. To je iznimno složen postupak jer podrazumijeva razumijevanje svojstava koja mogu igrati ulogu u traženju odgovora na istraživačko pitanje, što nije uvijek očito, a što se, kako autori navode, može ilustrirati činjenicom da se sintaktički učinci ne mogu razjasniti zanemarenim kontekstualnim svojstvima (usp. De Kuthy i Meurers, 2003). Korpusni podaci dobiveni traženjem lingvistički relevantnoga uzorka pokazuju široku varijaciju poznatih i nepoznatih parametara i mogu uključivati informacije o kontekstu, prema potrebi za istraživanje interakcije ograničenja iz sintakse i formalne pragmatike. Stoga je, prilikom traženja određenoga uzorka u korpusu, moguće uočiti teorijski zanimljiv obrazac unutar rečenica koje pokazuju široku varijaciju leksičkih, sintaktičkih, semantičkih i kontekstualnih svojstava, što omogućuje bolji uvid u to koja svojstva su relevantna za jezičnu pojavu koja se istražuje. Činjenica da su korpusni primjeri općenito prirodni i kontekstualizirani također može biti od pomoći kada se primjeri procjenjuju introspekcijom. Nadalje, valja imati na umu da gramatički priručnici koji nisu

utemeljeni na korpusnim istraživanjima katkada važnost pridaju jezičnome aspektu koji nije frekventan u stvarnoj jezičnoj uporabi, kao što je to primjerice *present continuous* u engleskome (Biber i Conrad, 2010). Posavec (2018: 71), oslanjajući se na Dashova (2011) promišljanja, pokazala je kako koristiti računalne korpuse za izradu nastavnih materijala i podučavanje hrvatskoga jezika. Autorica navodi da „za razliku od tradicionalnih udžbenika, koji se svojim primjerima jezične uporabe oslanjaju na stereotipne i ponekad suhoparne primjere te sadržavaju intuitivno izmišljene primjere i informacije koje općenito ignoriraju važne aspekte korištenja jezika, korpusi nude primjere korištenja jezika u njegovu svakodnevnom okruženju čime takvi primjeri postaju pouzdanijima i autentičnijima.“ Nadamo se da smo pokazali bogatstvo primjera koji inače ne bi bili vidljivi.

8. ZAKLJUČAK

Korpusna je lingvistika jedan od istraživačkih pristupa ispitivanju i proučavanju jezika. Jedna je od osnovnih značajki korpusnih metoda empirijsko ispitivanje jezičnoga ostvaraja koje se temelji na analizi velike (i uravnotežene) zbirke tekstova. Korpusna lingvistika koristi dostignuća računalne tehnologije te primjenjuje automatske i poluautomatske metode, a ovisi o kvantitativnim i kvalitativnim analitičkim metodama.

U knjizi su prikazana kvalitativna i kvantitativna istraživanja, a rezultati mogu doprinijeti novim spoznajama o ograničenjima korpusnih istraživanja na postojećim korpusima hrvatskoga jezika.

U ovoj su knjizi na primjeru evidencije, frekvencije i relacije prikazani prednosti i nedostaci korpusnih metoda u pronalaženju odgovora na istraživačka pitanja. Autorica je pokazala kako sastaviti vlastiti korpus te kritički promotriti dobivene rezultate, za što je potrebno razumijevanje (barem u nekoj mjeri) procesa koji su u pozadini računalnojezikoslovnih alata. Stoga je u knjigu uvrštena i osnovna terminologija koja čitatelju pomaže razlučiti osnovne pojmove i procese korpusom vođenih i korpusno utemeljenih istraživanja.

U prvome je poglavlju prikazan kratki povijesni pregled razvoja korpusne lingvistike s posebnim osvrtom na razvoj korpusne lingvistike i računalnojezikoslovnih alata za hrvatski jezik te se daje iscrpan i sustavan popis resursa i alata za hrvatski. U ovome se poglavlju također raspravlja o povezanosti korpusne lingvistike s drugim jezičnim i nejezičnim disciplinama te o ključnim pojmovima kao što su uravnoteženost, reprezentativnost i veličina korpusa.

U žarištu je drugoga poglavlja obilježavanje korpusa, što je od velike važnosti jer obilježeni korpus omogućuje ponavljanje korpusno utemeljenih ili korpusom vođenih istraživanja, čime postaje višefunkcionalnim jer se kao metodološki konstrukt može koristiti u leksikografiji, strojnom prevođenju, poučavanju jezika, analizi diskursa i mnogim drugim potpodručjima.

Treće poglavlje sadrži opis, pojašnjenje i ilustraciju osnovnih pojmova koji se rabe u korpusnoj lingvistici uz poseban osvrt na jednu vrstu programskoga jezika visoke razine i regularne izraze, bez razumijevanja kojih nije moguće adekvatno dohvatiti i interpretirati podatke iz korpusa.

U četvrtom su poglavlju opisani načini izrade korpusa, te osnovni pojmovi i procesi nužni pri izradi korpusa. Pokazani su prednosti i nedostaci alata za izradu korpusa za potrebe vlastitih istraživanja s posebnim osvrtom na korpus *ENGRI* koji je sastavljen u sklopu projekta *Engleske riječi u hrvatskome jeziku: identifikacija, afektivno-semantičko normiranje i ispitivanje kognitivne obrade bihevioralnim i neuroznanstvenim metodama* (HRZZ, 2020.-2025.) i koji, uz *Bazu engleskih riječi i hrvatskih istovrijednica,* nastalu na temelju korpusnih istraživanja, predstavlja jedan od recentnijih doprinosa resursima hrvatskoga jezika. Četvrto i peto poglavlje odnose se na dohvat podataka iz korpusa. U četvrtome su poglavlju u žarištu frekvencijske liste koje su polazište za daljnja korpusna istraživanja, prije svega u leksikografiji i poučavanju jezika. Ova dva poglavlja prikazuju vrste podataka koje možemo dobiti iz korpusa, odnosno jezičnoga ostvaraja, a to su popisi riječi (evidencije) s brojanjem (frekvencije). U šestome se poglavlju analiziraju načini crpljenja engleskih riječi iz korpusa hrvatskoga jezika. Sedmo i posljednje poglavlje kroz prizmu leksikogramatike prikazuje prednosti i nedostatke metoda korpusne lingvistike u provođenju istraživanja na razini riječi, skupine i (su)rečenice. Pri tome uvijek valja imati na umu da je kvantitativna analiza neodvojiva od kvalitativne analize.

Ostaje niz neistraženih tema kojima bi valjalo posvetiti pažnju, kao što su učenički korpusi i dijakronijski korpusi, koji do sada nisu bili u žarištu autoričina ispitivanja, te stoga nisu uvršteni u ovu knjigu.

I na kraju, važna jezikoslovna pitanja na koja trenutni računalnojezikoslovni resursi i alati ne mogu dati odgovor trebaju se i dalje sagledavati kroz prizmu korpusa, a upravo je suradnja jezikoslovaca, odnosno korisnika korpusa, i programera, koji razvijaju resurse i alate, neophodna za daljnji razvoj jezičnih tehnologija za hrvatski jezik.

POPIS LITERATURE

Agić, Ž., & Ljubešić, N. (2015). Universal dependencies for Croatian (that work for Serbian, too). U Piskorski, J, Pivovarova, L., Šnajder, J., Tanev, H, & Yangarber, R. (ur.), *The 5th Workshop on Balto-Slavic Natural Language Processing* (str. 1–8). Shoumen, Bugarska: INCOMA Ltd.

Agić, Ž., Tadić, M., & Dovedan, Z. (2009). Error analysis in Croatian morphosyntactic tagging. U Luzar-Stiffler, V., Jarec, I., & Bekic, Z. (ur.), *Proceedings of the International Conference on Information Technology Interfaces, ITI* (str. 521–526). https://doi.org/10.1109/ITI.2009.5196140.

Agić, Ž., Dovedan, Z., & Tadić, M. (2008). Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica, 32*(4), 445–451.

Agić, Ž., & Tadić, M. (2006). Evaluating morphosyntactic tagging of Croatian texts. U Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J. & Tapias, D (ur.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*. Genova, Italija: European Language Resources Association. Preuzeto s: http://www.lrec-conf.org/proceedings/lrec2006/pdf/326_pdf.pdf.

Alex, B. (2005). An unsupervised system for identifying English inclusions in German text. U Knight, K., Tou Ng, H., & Oflazer, K. (ur.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (str. 133–138). https://doi.org/10.3115/1628960.1628985.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of current word combinations. U Cowie, A. P. (ur.), *Phraseology: Theory, analysis and applications* (str. 101–122). Oxford: Oxford University Press.

Alvarez-Mellado, E. (2020). An annotated corpus of emerging anglicisms in Spanish newspaper headlines. U Thamar, S., Monojit, C, Kalika, B., Sunayana, S., Amitava, D., & Mona, D. (ur), *Proceedings of the 4th Workshop on Computational Approaches to Code Switching (LREC 2020)* (str. 1–8). Marseille: European Language Resources Association.

Anthony, L. (2019b). Resources for researching vocabulary. U Webb, S. (ur.), *The Routledge Handbook of Vocabulary Studies* (str. 561–590). Taylor and Francis. https://doi.org/10.4324/9780429291586-35.

Baker, P., & Egbert, J. (2016). *Triangulating Methodological Approaches in Corpus Linguistic Research*. London i New York: Routledge. https://doi.org/10.4324/9781315724812.

Balteiro, I. (2011). A reassessment of traditional lexicographical tools in the light of new corpora: sports Anglicisms in Spanish. *International Journal of English Studies,* 11(2), 23–52. https://doi.org/10.6018/ijes/2011/2/149631.

Bañón, M., M. ; Forcada, M. L. ; C. ; Kuzman, T. ; Ljubešić, N. ; van Noord, R. ; Pla Sempere, L. ; Ramírez-Sánchez, G.; Rupnik, P. ; Suchomel, V. ; Toral, A. ; van der Werff, T. ; Zaragoza, J. (2002). *Croatian web corpus MaCoCu-hr 1.0.* Slovenian language resource repository CLARIN.SI. Preuzeto s: http://hdl.handle.net/11356/1516.

Barlow, M. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics, 4*(1). https://doi.org/10.1075/ijcl.4.1.09bar.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. *12th EURALEX International Congress*. Preuzeto s: https://www.academia.edu/es/2713037/BootCaT_Bootstrapping_corpora_and_terms_from_the_web.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. U Lino, M.T., Xavier, M.F., Ferreira, F. Costa, R., & Silva, R. (ur.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (str. 1313–1316). Lisabon: European Language Resources Association.

Bekavac, B. (2001). *Primjena računalnojezikoslovnih alata na hrvatske korpuse.* Zagreb: Filozofski fakultet Sveučilišta u Zagrebu.

Bekavac, B. (2002). Strojno obilježavanje hrvatskih tekstova – stanje i perspektive. *Suvremena lingvistika*, *53-54* (1-2), 173–182.

Bekavac, B., & Tadić, M. (2003). Preparation of POS tagging of Croatian using CLaRK System. U Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., & Nikolov, N. (ur.), *Proceedings of RANLP 2003 Conference* (str. 455–459). Borovets, Bugarska.

Berman, R. A. (2008). Developing Linguistic Knowledge and Language Use Across Adolescence. U Hoff, E., & Shatz, M. (ur.), *Blackwell Handbook of Language Development* (str. 347–367). https://doi.org/10.1002/9780470757833.ch17.

Biber, D. (2015). Corpus-Based and Corpus-Driven Analyses of Language Variation and Use. U Heine, B., Narrod, H. & Biber, D. (ur.), *The Oxford Handbook of Linguistic Analysis* (str. 159–192). https://doi.org/10.1093/oxfordhb/9780199544004.013.0008.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib.

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers* (*Studies in Corpus Linguistics Issue 23*). Amsterdam i Philadelphia: John Benjamins.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243.

Biber, D., & Reppen, R. (2015). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1007/9781139764377.

Biber, D., & Conrad, S. (2010). Corpus Linguistics and Grammar Teaching. Preuzeto s: http://www.longmanhomeusa.com/content/pl_biber_conrad_monograph5_lo.pdf.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK: Pearson.

Bloomfield, L. (1933). *Language*. New York: Holt.

Bogetić, K. (2021). METALANGCORP: Presenting the First Corpus of Media Metalanguage in Slovene, Croatian and Serbian, and its Cross-discipline Applicability. *Fluminensia, 33*(1), 123–142.

Bogetić, K., & Batanović, V. (2020). *Annotated corpus of Croatian language-related news articles*. Slovenian language resource repository CLARIN.SI. Preuzeto s: http://hdl.handle.net/11356/1369.

Bogunović, I., Jelčić Čolakovac, J., & Borucinsky, M. (2022). Baza engleskih riječi i hrvatskih istovrijednica [Skup podataka]. Rijeka: Pomorski fakultet, Repozitorij. Dohvaćeno iz https://urn.nsk.hr/urn:nbn:hr:187:891063.

Bogunović, I., Kučić, M., Ljubešić, N., & Erjavec, T. (2021). *Korpus hrvatskih internetskih portala* (2014-2018). Preuzeto s: http://hdl.handle.net/11356/1416.

Bogunović, I., & Ćoso, B. (2013). Engleski u hrvatskome: znanstveni izričaj biomedicine i zdravstva. *Fluminensia*, 25(2), 177–191.

Bogunović, I., & Kučić, M. (u postupku recenzije). Classification of English words in Croatian: A comparison between hand-engineered and N-gram features.

Borucinsky, M. (2017). Korpusansätze in der Sprachforschung: Mit besonderer Rücksicht auf korpusgebundene Untersuchungen der kroatischen Sprache. U Cergol Kovačević, K., & Udier, S.L. (ur.), *Applied Linguistics Research and Methodology – Proceedings from the 2015 CALS conference* (str. 255–269). Frankfurt am Main: Peter Lang.

Borucinsky, M. (2015). *Modifikacija u imenskim skupinama u engleskome i hrvatskome jeziku*. Doktorska disertacija. Zagreb: Filozofski fakultet.

Borucinsky, M. & Pritchard, B. (2022). Lexical Bundles in Maritime Texts. *ICAME Journal: Computers in English Linguistics*, 46(1), 5–17. https://doi:10.2478/icame-2022-0001.

Borucinsky, M., & Bogunović, I. (2022). Crpljenje engleskih riječi iz korpusa hrvatskoga jezika. *Fluminensia,* 34(2), 435–461.

Borucinsky, M., & Tominac Coslovich, S. (2021). Introducing data-driven learning into the marine engineering English classroom. *Humanities Science Current Issues*, *3*(42), 19–27. https://doi.org/10.24919/2308-4863/42-3-4.

Borucinsky, M., & Kegalj, J. (2019). Syntactic ambiguity of (complex) nominal groups in technical English. *International Journal of English Studies*, 19(2), 83–102. https://doi.org/10.6018/IJES.352751.

Borucinsky, M., & Kegalj, J. (2017). Višerječni nazivi u jeziku brodostrojarske struke. U Omrčen, D., & Krakić, A.-M. (ur.), *Od teorije do prakse u jeziku struke* (str. 7–23), Zagreb: Udruga nastavnika jezika struke.

Boulton, A. (2015). Applying data-driven learning to the web. U Leńko-Szymańska, A. & Boulton, A. (ur.), *Multiple Affordances of Language Corpora for Data-driven Learning* (str. 267–296). ATILF, CNRS & University of Lorraine. https://doi.org/10.1075/scl.69.13bou.

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language*. London: Routledge. https://doi.org/10.4324/9780203469255

Bratanić, M. (1998). Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. *Filologija*, 30-31, 171–177.

Brdar, I. (2010). Engleske riječi u jeziku hrvatskih medija. *Lahor*, *2*(10), 217–232.

Březina, V. (2021). *A drop in the ocean: Corpora as samples of language.* Lancaster: Lancaster Corpus Linguistics. Preuzeto s: https://www.futurelearn.com/info/courses/corpus-linguistics/0/steps/261921.

Březina, V. (2018). *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Březina , V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre.

Březina, V., & Pořízka, P. (2021). Collocation graphs and networks using #lancsbox: Applications in English and Czech. *Casopis pro Moderni Filologii*, *103*(1), 36–59. https://doi.org/10.14712/23366591.2021.1.2.

Brozović, D., Čavar, D. i T. Erjavec (2018). *Croatian language corpus Riznica*. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1180.

Carter, R., & McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Castro, D. W., Souza, E., Vitório, D., Santos, D., & Oliveira, A. L. I. (2017). Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties. *Applied Soft Computing Journal*, *61*, 1160–1172. https://doi.org/10.1016/j.asoc.2017.05.065.

Cheng, W., Greaves, C., Sinclair, J. M. H., & Warren, M. (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*, *30*(2), 236–252. https://doi.org/10.1093/applin/amn039.

Cheng, W. L., Greaves, C., & Warren, M. J. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*, 411–433.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Computational Linguistuics 16,* 76–83. https://doi.org/10.3115/981623.981633.

Corpas, G., & Seghiri, M. (2009). Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish). In A. Beeby, P. Rodríguez, & P. Sánchez-Gijón (ur.), *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate* (str. 75–107). Amsterdam, NLD: John Benjamins.

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2) , 94–100. https://doi.org/10.1080/09296171003643098.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966.

Ćoso, B., & Bogunović, I. (2017). Person perception and language: A case of English words in Croatian. *Language and Communication*, 53, 25–34. https://doi.org/10.1016/j.langcom.2016.11.001.

Dash, N. S. (2011) *Corpus-Based English Language Teaching: A New Method*. Preuzeto s: https://www.academia.edu/3563414/Corpus-Based_English_Language_Teaching_A_New_Method.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures*, 225–246.

De Kuthy, K. i W. D. Meurers. (2003). Dealing with Optional Complements in HPSG-Based Grammar Implementations. U Müller, S. (ur.), *Proceedings of the HPSG03 Conference.* Michigan State University, East Lansing. Preuzeto s: http://csli-publications.stanford.edu/

Delecraz, S., Becerra-Bonache, L., Favre, B., Nasr, A., & Bechet, F. (2021). Multimodal Machine Learning for Natural Language Processing: Disambiguating Prepositional Phrase Attachments with Images. *Neural Processing Letters*, *53*(5), 309–3121. https://doi.org/10.1007/s11063-020-10314-8.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297–302. doi:doi10.2307/1932409.

Dickinson, M., & Meurers, W. D. (2005). Detecting errors in discontinuous structural annotation. U Knight, K., Ng, H. T. & Oflazer, K. (ur.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)* (str. 322–329). Michigan: Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219880.

Dickinson, M., & Meurers, W. D. (2003). Detecting errors in part-of-speech annotation. U *10th Conference of the European Chapter of the Association for Computational Linguistics*, *EACL 2003* (str. 107-114). Budapest, Hungary. Preuzeto s: https://www.sfs.uni-tuebingen.de/~dm/papers/dickinson-meurers-03.pdf. https://doi.org/10.3115/1067807.1067823.

Duvnjak Jardas, I. (2019). Anglicizmi u sportskoj terminologiji u hrvatskom jeziku. *Zbornik radova Veleučilišta u Šibeniku*, 185–194. Preuzeto s: https://hrcak.srce.hr/223022.

Ebeling, J., Oksefjell Ebeling, S., & Hasselgård, H. (2013). Using recurrent word-combinations to explore cross-linguistic differences. U Aijmer, K. & Altenberg, B. (ur.), *Advances in Corpus-based Contrastive Linguistics: Studies in honour of Stig Johansson* (str. 177–200). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.54.11ebe.

Erjavec, T. (2021). M*ultilingual comparable corpora of parliamentary debates ParlaMint 2.1*. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1432.

Evert, S. (2005). *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Doktorska disertacija. Stuttgart: Institut Für Maschinelle Sprachverarbeitung.

Fandrych, Ch. (2006). Bildhaftigkeit und Formelhaftigkeit in der allgemeinen Wissenschaftssprache als Herausforderung für Deutsch als Fremdsprache. U Ehlich, K. & Heller, E. (ur.), *Die Wissenschaft und ihre Sprachen* (str. 39–62). Frankfurt am Main: Peter Lang.

Fellbaum, C. (2014). Large-scale lexicography in the digital age. *International Journal of Lexicography, 27*(4), 378–395. https://doi.org/10.1093/ijl/ecu018.

Filipović Petrović, I. (2020). Kako Iggyju reći pop, a Dylanu bob: iz korpusnih istraživanja frazema. *Filologija : Časopis Razreda Za Filološke Znanosti Hrvatske Akademije Znanosti i Umjetnosti,* 75, 147–164. https://doi.org/10.21857/ydkx2cwxk9.

Filipović Petrović, I. (2018). *Kada se sretnu leksikografija i frazeologija. O statusu frazema u rječniku*. Zagreb: Srednja Europa.

Filipović Petrović, I., & Parizoska, J. (2019). Konceptualna organizacija frazeoloških rječnika u e-leksikografiji. *Filologija : Časopis Razreda Za Filološke Znanosti Hrvatske Akademije Znanosti i Umjetnosti*, 73, 27–45. https://doi.org/10.21857/moxpjhgwpm.

Filipović, R. (1990) *Anglicizmi u hrvatskom ili srpskom jeziku: porijeklo-razvoj-značenje*. Zagreb: Školska knjiga.

Filko, M. (2020). *Unutarleksičke i međuleksičke strukture imeničkoga dijela hrvatskoga leksika. Doktorska disertacija*. Zagreb: Filozofski fakultet.

Fillmore, C. (1992). Corpus linguistics" vs. "computer-aided armchair linguistics. *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics* (str. 35–66).

Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-55. In Studies in Linguistic Analysis (str. 1–31). *Special Volume of the Philological Society*. Oxford: Blackwell.

Fletcher, W. H. (2012). Corpus Analysis of the World Wide Web. U Chapelle, C. A. (ur.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0254.

Forta, B. (2018). *Learning regular expressions*. Boston: Addison-Wesley Professional.

Furiassi, C. (2008). What dictionaries leave out: new non-adapted anglicisms in Italian. U A. Martelli & Pulcini, V. (ur.), *Investigating English with corpora. Studies in honor of Maria Teresa Prat* (str. 153–169). Polimetrica International Scientific Publisher.

Furiassi, C. & Hofland, K. (2007). The retrieval of false anglicisms in newspaper texts. U Facchinetti, R. (ur.), *Corpus Linguistics 25 Years on. Language and Computers 62, Studies in Practical Linguistics* (str. 347–363.). Amsterdam i New York: Rodopi.

Gablasova, D., Březina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning* 67, 155–179. https://doi.org/10.1111/lang.12225.

Garside, R. G., Leech, G., & McEnery, A. M. (2020) (ur.). *Corpus Annotation. Linguistic Information from Computer Text Corpora.* London/New York: Routledge. https://doi.org/10.4324/9781315841366-7.

Gray, B. & Biber, D. (2015). Phraseology. U Biber, D. & Reppen, R. (ur.), *The Cambridge Handbook of English Corpus Linguistics* (str. 125–145). Cambridge: Cambridge University Press.

Gries, S. Th. (2012). Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics; Towards more and more fruitful changes. U Mukherjee, J. & Huber, M. (ur.), *Corpus Linguistics and Variation in English* (str. 41-63).https://doi.org/10.1163/9789401207713_006. Preuzeto s: https://www.researchgate.net/publication/263299861_Corpus_linguistics_theoretical_linguistics_and_cognitivepsycholinguistics_Towards_more_and_more_fruitful_exchanges.

Gries, S. Th. (2010). Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily... *International Journal of Corpus Linguistics, 15*(3), 327–343. https://doi.org/10.1075/ijcl.15.3.02gri.

Gries, S. Th. (2009). *Quantitative corpus linguistics with R: A practical introduction*. Taylor and Francis. https://doi.org/10.4324/9780203880920.

Hall, R., Wittgenstein, L., & Anscombe, G. E. M. (1967). Philosophical Investigations. *The Philosophical Quarterly, 17(*69), 362–363. https://doi.org/10.2307/2217461.

Halliday, M. A. K. (1996). On Grammar and Grammatics. U Hasan, R. Cloran, C. & Butt, D. G. (ur.). *Functional Descriptions. Theory in Practice* (str. 1–38). Amsterdam: John Benjamins.

Halliday, M. A. K. (1961). Categories of the Theory of Grammar. *Word* 17, 241–292. Pretisak u Webster, J. (2002) (ur.) *The Collected Works of M. A. K. Halliday. On Grammar, Vol*. I, (str. 37–95). London i New York: Continuum.

Halliday, M., & Matthiessen, C. (2014). *An Introduction to Functional Grammar*. New York: Routledge. https://doi.org/10.4324/9780203783771.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), 398–436. https://doi.org/10.1093/ijl/ecs026.

Hansack, E., Hansen, B. W., Horvat, M., & Perić Gavrančić, S. (2016). Regenburški dijakronijski korpus hrvatskoga jezika - CRODI. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 52*(1), 1–19.

Hardie, A., & McEnery, T. (2010). On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics, 15*(3), 384–394. https://doi.org/10.1075/ijcl.15.3.09har.

Harris, Z. S. (1954). Distributional Structure. *WORD*, 10, (2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520.

Hasanić, S. (2017). Alati i tražilice za pretraživanje korpusa. Diplomski rad. Zagreb: Sveučilište u Zagrebu.

Hausmann, F. J. (2007). Die Kollokationen im Rahmen der Phraseologie - Systematische und historische Darstellung. *Zeitschrift für Anglistik und Amerikanistik*, *55*(3), 217–234. https://doi.org/10.1515/zaa.2007.55.3.217.

Hlaváčová, J. (2006). New approach to frequency dictionaries - Czech example. U Nicoletta Calzolari, N., Choukri, K., Gangemi,A., Maegaard, B., Mariani, J., Odijk, J. & Tapias, D. (ur.), *Proceedings of the 5th International Conference on Language Resources and Evaluation* (str. 373–378). European Language Resources Association. Preuzeto s: http://www.lrec-conf.org/proceedings/lrec2006/pdf/11_pdf.pdf.

Hržica, G., Košutar, S., & Posavec, K. (2021). Connectives and other discourse markers in written language and spontaneous speech. *Fluminensia*, 33(1), 25–52. https://doi.org/10.31820/F.33.1.12.

Hughes, B., Baldwin, T., Bird, S., Nicholson, J., & MacKinlay, A. (2006). Reconsidering language identification for written language resources. U Nicoletta Calzolari, N., Choukri, K., Gangemi,A., Maegaard, B., Mariani, J., Odijk, J. & Tapias, D. (ur.), *Proceedings of the 5th International Conference on Language Resources and Evaluation* (str. 485-488). European Language Resources Association. Preuzeto s: http://www.lrec-conf.org/proceedings/lrec2006/pdf/459_pdf.pdf.

Hundt, M., Nesselhauf, N., & Biewer, C. (2015). Corpus linguistics and the web. Amsterdam i Atlanta: Rodopi. https://doi.org/10.1163/9789401203791_002.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. & Francis, G. (2002). *Pattern Grammar. A Corpus-driven Approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001.

Jakubiček, M., Kovář, V., Rychlý, P., & Suchomel, V. (2020). Current Challenges in Web Corpus Building. U Jakubíček, M., Kovář, V., Rychlý, P. & Suchomel, V. (ur.), *Proceedings of the 12th Web as Corpus Workshop* (str. 1–4). Marseille, Francuska. Preuzeto s: https://aclanthology.org/2020.wac-1.1.pdf.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research,* 65, 675–782. https://doi.org/10.1613/JAIR.1.11675.

Jelaska, Z., & Baričević, V. (2012). Leksička jednostavnost i značenjska složenost rječnika Ivanova evanđelja. *Lahor*, 13, 102–137.

Jelčić Čolakovac, J. & Borucinsky, M. (u pripremi). In the Melting Pot of Web-Crawled Texts: The Challenges of Extracting English Words and Phrases.

Johansson, V. (2008). Lexical diversity and Lexical Density In Speech and Writing: A Developmental Perspective. *Working Papers in Linguistics, 53*, 61–79.

Johnston, J. E., Berry, K. J., & Mielke, P. W. (2006). Measures of effect size for chi-squared and likelihood-ratio goodness-of-fit tests. *Perceptual and Motor Skills, 103*(2), 412–414. https://doi.org/10.2466/PMS.103.2.412-414.

Jones, S., Paradis, C., Murphy, M. L., & Willners, C. (2007). Googling for 'opposites': a web-based study of antonym canonicity. *Corpora, 2*(2), 129-155. https://doi.org/10.3366/cor.2007.2.2.129.

Kegalj, J. Žanrovska analiza pomorskopravnih tekstova i ostvarenje prijevodnih univerzalija u njihovim prijevodima s engleskoga jezika. Doktorska disertacija (u pripremi). Zagreb: Filozofski fakultet Sveučilišta u Zagrebu.

Kegalj, J., & Borucinsky, M. (2022). Genre-based approach to corpus compilation for translation research. U Petkova, T. V. & Chukov, V. S. (ur.). *7th International e-Conference on Studies in Humanities and Social Sciences: Conference Proceedings* (str. 215–224). Beograd: Center for Open Access in Science. https://doi.org/10.32591/coas.e-conf.07.22215k.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., & Rychlý, P. (2015). *Statistics used in the Sketch Engine*. Preuzeto s: https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/.

Kilgarriff, A., & Kosem, I. (2012). Corpus tools for lexicographers. U Granger, S. & Paquot, M. (ur.), *Electronic Lexicography*. Oxford: Oxford Academic web. https://doi.org/10.1093/acprof:oso/9780199654864.003.0003.

Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK. Preuzeto s: https://www.coursehero.com/file/100979453/2009-Simple-maths-for-keywordspdf/.

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. Preuzeto s: https://www.euralex.org/elx_proceedings/Euralex2008/095_Euralex_2008_Adam%20Kilgarriff_Milos%20Husak_Katy%20McAdam_Michael%20Rundell_Pavel%20Rychly_GDEX_Automatically%20Finding%20Good%20Di.pdf.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), 263–276. https://doi.org/10.1515/CLLT.2005.1.2.263.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2004). The Sketch Engine: ten years on. *Lexicography*, 1, 7–36.

Kilgarriff, A. & Grefenstette, G. (2003). *Introduction to the Special Issue on the Web as Corpus*. Preuzeto s: https://www.academia.edu/1541459/Introduction_to_the_Special_Issue_on_the_Web_as_Corpus.

Kilgarriff, A., & Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. *Proceedings of the ACL Workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation* (str. 32–38). Toulouse: Association for Computational Linguistics.

Knight, D. (2011). The future of multimodal corpora. *Brazilian Journal of Applied Linguistics* (BJAS), 11(2), 391–415.

Knight, D., Evans, D., Carter, R. A., & Adolphs, S. (2009). Redrafting corpus development methodologies: Blueprints for 3rd generation multimodal, multimedia corpora. *Corpora,* 4(1), 1–32.

Kolarović, V., & Tadić, M. (2013). *Korpus (katalog izložbe) Knjižnice grada Zagreba*. Zagreb: Galerija VN Knjižnice Vladimira Nazora.

Kovačević, M. (2002). *Croatian corpus, CHILDES*. Preuzeto s: http://childes.psy.cmu.edu/data/Slavic

Kraljević, J. K., & Hržica, G. (2017). Croatian adult spoken language corpus (HrAL). *Fluminensia*, 28 (2), 87–102.

Kucera, Henry, & Francis, W. N. (1970). *Computational analysis of present-day American English*. Providence: Brown Unviersity press.

Kučić, M. (2021). Creating a Web Corpus Using GO. U Skala, K. (ur.), *44. Proceedings of the International Convention MIPRO* (str. 1931–1934). Rijeka: Institute of Electrical and Electronics Engineers. https://doi.org/10.23919/MIPRO52101.2021.9597093.

Kuvač Kraljević, J., Hržica, G., Štefanec, V., Kologranić Belić, L., & Ljubešić, N. (2021). Croatian corpus of non-professional written language by typical speakers and speakers with language disorders RAPUT 1.0. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1435.

Kuvač Kraljević, J., Hržića, G., Olujić, M., Kologranić Belić, L., Palmović, M., & Matić, A. (2016). Uzorkovanje specijaliziranih govornih i pisanih korpusa jezika odraslih govornika: izazovi i nedoumice. U Cergol Kovačević, K. & Udier, S. L. (ur.), *Metodologija i primjena lingvističkih istraživanja, Zbornik radova HDPL*-a, (str. 159–170). Zagreb: Hrvatsko društvo za primijenjenu lingvistiku.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, 9(9) e105825 https://doi.org/10.1371/JOURNAL.PONE.0105825.

Květoň, P., & Oliva, K. (2002). Achieving an almost correct PoS-tagged corpus. U Sojka, P., Kopeček, I. & Pala, K. (ur.), *International Conference on Text, Speech and Dialogue*. Part of the Lecture Notes in Computer Science book series (LNAI, 2448). Berlin i Heidenberg: Springer. https://doi.org/10.1007/3-540-46154-x_3.

Lalli Paćelat, I. (2014). *Analiza zakonopravnoga stila hrvatskoga i talijanskoga jezika: unutarjezična, međujezična i prijevodna perspektiva*. Doktorska disertacija. Zagreb: Sveučilište u Zagrebu.

Leech, G. N. (2011). Frequency, corpora and language learning. U Meunier, F., DeCock, S. Gilquin, G. & Granger, S. (ur.), *A Taste for Corpora: In honour of Sylviane Granger* (str. 7–32). Amsterdam: John Benjamins. https://doi.org/10.1075/SCL.45.05LEE.

Leech, G. N. (1992). *Corpora and Theories of Linguistic Performance*. Berlin: Mouton de Gruyter.

Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature*, 520(br. 7549), 612. https://doi.org/10.1038/520612A.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31(2), 374–397. https://doi.org/10.1093/LLC/FQU064.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburg: Edinburgh University Press. https://doi.org/10.2478/icame-2020-0006.

Losnegaard, G. S., & Lyse, G. I. (2012). A data-driven approach to anglicism identification in Norwegian. U Andersen, G. (ur.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian* (str. 131–154). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.49.07los.

Ljubešić, N., & Lauc, D. (2021). BERTić - The transformer language model for Bosnian, Croatian, Montenegrin and Serbian. U Babych, B. i dr. (ur.), *Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, BSNLP* (str. 37–42). Kiyv, Ukrajina: Association for Computational Linguistics.

Ljubešić, N., Markoski, F., Markoska, E., & Erjavec, T. (2021). *Comparable corpora of South-Slavic Wikipedias*. Slovenian language resource repository. Preuzeto s: https://www.clarin.si/repository/xmlui/handle/11356/1427.

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. U Erjavec, T., Marcińczuk, M., Nakov, P., Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J. & Yangarber, R.(ur.), *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (str. 29–34). Firenza, Italija.

Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., & Samardžić, T. (2019). *Croatian Twitter training corpus* ReLDI-NormTagNER-hr 2.1. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1241.

Ljubešić, N., Agić, Ž., Klubička, F., Batanović, V., & Erjavec, T. (2018). *Training corpus hr500k* 1.0. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1183.

Ljubešić, N., Agić, Ž., Klubička, F., Batanović, V., & Erjavec, T. (2018). hr500k – A Reference Training Corpus of Croatian. *Proceedings of the Conference on Language Technologies & Digital Humanities* 2018 (JT-DH 2018) (str. 154–161). Ljubljana. Preuzeto s: https://vukbatanovic.github.io/pdf/JTDH_HR_2018.pdf.

Ljubešić, N., Esplà-Gomis, M., Ortiz Rojas, S., Klubička, F., & Toral, A. (2016). *Croatian-English parallel corpus hrenWaC 2.0.* Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1058.

Ljubešić, N., & Erjavec, T. (2016). U Calzolari, N., Choukri, K., Thierry, D., Goggi, S., Grobelnik, M. & Maegaard, B. (ur.), *Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016) (str. 1527–1531). Pariz: European Language Resources Association (ELRA).

Ljubešić, N., Fišer, D., & Erjavec, T. (2014). TweetCaT: A tool for building Twitter corpora of smaller languages. U Calzolari, N., Choukri, K., Declerck, Th., Loftsson, H., Maegaard, B., Mariani, J. Moreno, A., Odijk, J. & Piperidis, S. (ur.), *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC* (str. 2279–2283). Reykjavik, Island: European Language Resources Association (ELRA).

Ljubešić, N., & Klubička, F. (2014). {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. U Bildhauer, F. & Schäfer, A. (ur.). *Proceedings of the 9th Web as Corpus Workshop* (str. 29–35). Gothenburg, Švedska: Association for ComputationalLinguistics. https://doi.org/10.3115/v1/w14-0405.

Ljubešić, N., & Erjavec, T. (2011). hrWaC and slWac: Compiling web corpora for Croatian and Slovene. U Habernal, I. & Matoušek, V. (ur.), *Lecture Notes in Computer Science* (str. 395–402). Berlin i Heidelberg: Springer.https://doi.org/10.1007/978-3-642-23538-2_50.

McCarthy, M. (2004). Using Corpora in Language Teaching. *CALPER Digests*. Penn State.

McEnery, T. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

McEnery, T., Březina, V., Gablasova, D., & Banerjee, J. (2019). Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics* 39, 74–92. https://doi.org/10.1017/S0267190519000096.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Cambridge textbooks in linguistics Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, A., Xiao, Z., & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London/New York: Routledge.

McHardy Sinclair, J. & Mauranen, A. (2006). *Linear Unit Grammar: Integrating Speech and Writing.* Amsterdam: John Benjamins.

Međeral, K. (2016). Jezične bakterije – pomagači ili štetočine u jezičnome organizmu? *Hrvatski jezik 3*, 1–10.

Meurers, W. D., & Müller, S. (2009). Corpora and syntax. U Lüdeling, A. & Kytö, M. (ur.), *Corpus Linguistics: An International Handbook* Vol. 2. (str. 920–933). Berlin: de Gruyter.

Měchura, M. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. *Proceedings of eLex 2017 conference.* Leiden, Nizozemska. Preuzeto s: https://elex.link/elex2017/wp-content/uploads/2017/09/paper41.pdf.

Mikelenić, B. (2020). *Kontrastivna korpusna analiza prijedložne dopune u španjolskome i njezinih ekvivalenata u hrvatskome.* Doktorska disertacija. Zagreb: Sveučilište u Zagrebu.

Mikelić Preradović, N., Berać, M., & Boras, D. (2015). Learner Corpus of Croatian as a Second Language. U Cergol Kovačević, K.. &. Udie, S. L. (ur.), *Multidisciplinary Approaches to Multilingualism* (str. 107–126). Frankfurt am Main: Peter Lang.

Moguš, M., Bratanić, M., & Tadić, M. (1999). *Hrvatski čestotni rječnik*. Zagreb: Školska knjiga.

Mozetič, I., Grčar, M., & Smailović, J. (2020). Twitter sentiment for 15 European languages. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1054.

Muhvić-Dimanovski, V. (2005). *Neologizmi: problemi teorije i primjene.* Zagreb: Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Muhvić-Dimanovski, V., Skelin Horvat, A. & Hriberski, D. (2016). *Rječnik neologizama u hrvatskome jeziku*. Preuzeto s: www.rjecnik.neologizam.ffzg.unizg.hr.

Mukherjee, J. (2010). Corpus linguistics versus corpus dogmatism — pace Wolfgang Teubert. *International Journal of Corpus Linguistics*, 15(3), 370–378. https://doi.org/10.1075/ijcl.15.3.07muk.

Nedić, I. (2017). *Model izrade učeničkog korpusa*. Diplomski rad. Zagreb: Filozofski fakultet, Sveučilište u Zagrebu.

Nesselhauf, N. (2005) *Corpus Linguistics: A Practical Introduction.* Preuzeto s: http://www.as.uniheidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Intro duction.pdf.

Núñez Nogueroles, E. E. (2016). Anglicisms in CREA: A Quantitative Analysis in Spanish Newspapers. *Language Design, 18,* 215–242.

Pandžić, I. (2015). Oblikovanje korjenovatelja za hrvatski jezik. *Raprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 41*(2), 301–327.

Pedersen, T. (1998). Dependent bigram identification. U Anon, A. (ur.), *Proceedings of the 1998 10th Conference on Innovative Applications of Artificial Intelligence, IAAI.*

Posavec, K. (2017). *Uloga računalnih korpusa u poučavanju hrvatskoga kao drugoga i inoga jezika*. Doktorska disertacija. Zagreb: Filozofski fakultet, Sveučilište u Zagrebu.

Posavec, K. (2018). Uporaba korpusa u poučavanju hrvatskoga kao drugoga i inoga jezika. *Studia lexicographica: časopis za leksikografiju i enciklopedistiku,* 12 (2) 63–84.

Precht, K., Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. *TESOL Quarterly*, 32(4), 789–790. https://doi.org/10.2307/3588017.

Pritchard, B. (1998). O kolokacijskom potencijalu rječničkog korpusa. *Filologija* 30 – 31, 285–304.

Purnelle, G., Fairon, C., & Dister, Anne. (2004) (ur.). *Le poids des mots : actes des 7es journées internationales d'analyse statistique des données textuelles. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*: Louvain-la-Neuve.

Purver, M., Shekhar, R., Pranjić, M., Pollak, S., & Martinc, M. (n.d.). 24sata news article archive 1.0,. 2021: Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1410.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. U *The Workshop on Comparing Corpora* (str. 1–6). Hong Kong, China. Association for Computational Linguistics. Preuzeto s: https://aclanthology.org/W00-0901.pdf. https://doi.org/10.3115/1117729.1117730.

Runjić-Stoilova, A. & A. Pandža (2010). Prilagodba anglizama u govoru na hrvatskim televizijama. *Croatian Studies Review, 6* (1), 229–240.

Rychlý, P. (2008). A lexicographer-friendly association score. RASLAN 2008 - Recent Advances in Slavonic Natural Language Processing: 2nd Workshop on Recent Advances in Slavonic Natural Language Processing, Proceedings.

Samardžić, T., Ljubešić, N., & Miličević, M. (2015). Regional Linguistic Data Initiative (ReLDI). U Piskorski, J., Pivovarova, L., Šnajder, J., Tanev, H. & Yangarber, R. (ur.), *The 5th Workshop on Balto-Slavic Natural Language Processing* (str. 40–42). INCOMA Ltd. Shoumen, Bugarska. Preuzeto s: https://aclanthology.org/W15-5306.pdf.

Samardžić, T., Starović, M., Agić, Ž., & Ljubešić, N. (2017). Universal dependencies for Serbian in comparison with Croatian and other Slavic Languages. U Erjavec, T., Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger,J. & Yangarber, R. (ur.), *6th Workshop on Balto-Slavic Natural*

Language Processing at the 15th Conference of the European Chapter of the Association for Computational Linguistics (str. 39–44). Valencia, Španjolska: Association for Computational Linguistics. Preuzeto s: https://aclanthology.org/W17-1407.pdf. https://doi.org/10.18653/v1/w17-1407.

Savický, P., & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, 9(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124.

Serigos, J. (2017). *Applying corpus and computational methods to loanword research: New approaches to anglicisms in Spanish*. Doktorska disertacija. Austin: The University of Texas at Austin.

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

Sinclair, J. (2004). Corpus and Text — Basic Principles. In *Developing Linguistic Corpora: a Guide to Good Practice*.

Stefanowitsch, A. (2020). *Corpus linguistics: A Guide to the methodology*. Berlin: Language Science Press.

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste.

Stubbs, M. (1993). British traditions in text analysis. U Baker, M. (ur.), *Text and Technology: In Honour of John Sinclair* (str. 1–33). Amsterdam: John Benjamins.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Massachusetts: Blackwell Publishers.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators. *Functions of Language*, 10(1), 61–104. https://doi.org/10.1075/fol.10.1.04stu.

Szudarski, P. (2017). *Corpus Linguistics for Vocabulary.* London: Routledge. https://doi.org/10.4324/9781315107769.

Šnjarić, M., & Borucinsky, M. (2020). Glagolsko-imeničke kolokacije hrvatskoga, njemačkoga i engleskoga općeznanstvenog jezika u općoj dvojezičnoj e-leksikografiji. *Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje,* 46(2), 1105–1127. https://doi.org/10.31724/rihjj.46.2.34.

Štefanec, V., Ljubešić, N., & Kraljević, J. K. (2016). Croatian error-annotated corpus of non-professional written language. U Calzolari, N., Choukri, K.,

Declerck, Th., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, M., Odijk, J. & Stelios Piperidis, S. (ur.), *Proceedings of the 10th International Conference on Language Resources and Evaluation* (str. 3220–3226). Portorož, Slovenija: European Language Resources Association (ELRA). Preuzeto s: https://aclanthology.org/L16-1513.pdf.

Štrkalj Despot, K., & Ostroški Anić. (2020). Pregled razvoja hrvatske e-leksikografije. U Brač, I. & Ostroški Anić, A. (ur.), *Svijet od riječi: terminološki i leksikografski ogledi*. (str. 5–24). Zagreb: Institut za hrvatski jezik i jezik i jezikoslovlje.

Tadić, M. (1997). Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. *Suvremena lingvistika, 23* (43/44), 387–394.

Tadić, M. (2001). Procedures in Building the Croatian-English Parallel Corpus. *International Journal of Corpus Linguistics, 6*(3), 107–123. https://doi.org/10.1075/ijcl.6.si.10tad.

Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik*. Zagreb Exlibris.

Tadić, M. (2009). New version of the Croatian National Corpus. U Hlaváčková, D. A. (ur.), *After Half a Century of Slavonic Natura*l *Language Processing* (str. 199–205). Brno, Republika Češka: Masaryk University.

Tadić, M., & Brozović-Rončević, D., & Kapetanović, A: (2012). *Hrvatski jezik u digitalnom dobu*. Heidelberg: Springer. doi:http://doi.org/10.1007/978-3-642-30882-6

Teubert, W. (1995). Language Resources: The Foundations of a Pan-European Information Society. *Proceedings of the First Trans-European Language Resource Infrastructure* (str. 105–128). Mannheim: Institut für deustche Sprache.

Teubert, W. (2005). My Version of Corpus Linguistics. *International Journal of Corpus Linguistics,* 10(1), 1–13.

Teubert, W. & Čermáková, A.. (2007). *Corpus linguistics: a short introduction*. London: Continuum.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6.

Tognini-Bonelli, E., & Sinclair, J. (2006). Corpora. U Brown, K. (ur.), *Encyclopedia of Language and Linguistics* (str. 206–219). Amsterdam: Elsevier.

Tomašić, A., & Brkić, M. (2012). Udio vrsta riječi u sva četiri evanđelja. *Lahor*, 14, 85–101.

Toral, A. e. (2016). Tourism English-Croatian Parallel Corpus 2.0,. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1049.

Tummers, J., Kris, H. & Geeraerts, D. (2005). Usage-based approaches in cognitive linguistics: a technical state of the art. *Corpus Linguistics and Linguistic Theory*, 1(2), 225–261.

Vasić, D., & al., e. (2020). Semantic hypergraph corpus SemCRO 1.0. Slovenian language resource repository. Preuzeto s: http://hdl.handle.net/11356/1377.

Wallis, S. (2020). *Statistics in Corpus Linguistics Research*. New York: Routledge. https://doi.org/10.4324/9780429491696.

Xiao, R. (2015). Collocation. U Biber, D. & Reppen, A. (ur.), *The Cambridge Handbook of English Corpus Linguistics* (str. 106–124). Cambridge: Cambridge University Press. doi:https://doi.org/10.1017/CBO9781139764377.007.

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics, 27*(1), 103–129. https://doi.org/10.1093/applin/ami045.

Zanettin, F. (2014). *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. London: Routledge. https:doi.org/10.4324/9781315759661.

Mrežni izvori, računalnojezikoslovni resursi i alati

Anthony, L. (2019a). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Preuzeto s: http://www.antlab.sci.waseda.ac.jp/.

Athelstan. (2010). ParaConc. Multilingual Concordancer. Preuzeto s: https://paraconc.com.

BNC | Lancaster University. Preuzeto s: http://corpora.lancs.ac.uk/bnclab/search.

Březina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Lancaster University.

British National Corpus. Preuzeto s: http://www.natcorp.ox.ac.uk/.

Build parallel and multilingual corpora | Sketch Engine. Preuzeto s: https://www.sketchengine.eu/guide/setting-up-parallel-corpora/#tab-id-2.

CLARIN ERIC. Preuzeto s: https://www.clarin.eu/.

CollTerm | Natural Language Processing group. Preuzeto s: http://nlp.ffzg.hr/resources/tools/collterm/.

CoRD | The Lancaster-Oslo/Bergen Corpus (LOB). Preuzeto s: https://varieng.helsinki.fi/CoRD/corpora/LOB/.

Corpus Analysis Tool. Preuzeto s: https://corpus-analysis.com/.

CQPweb Main Page. Preuzeto s: https://cqpweb.lancs.ac.uk/.

Dictionary.com | Meanings and Definitions of Words at Dictionary.com. Preuzeto s: https://www.dictionary.com/.

Digitales Wörterbuch der deutschen Sprache. Preuzeto s: https://www.dwds.de/.

EAGLES. Preuzeto s: http://www.ilc.cnr.it/EAGLES/home.html.

Fletcher, W. H. (2010). Phrases in English. Preuzeto s: http://phrasesinenglish.org/.

Gigafida Corpus - CJVT. Preuzeto s: https://www.cjvt.si/en/research/cjvt-projects/gigafida-corpus/.

GitHub - clarinsi/classla: CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages. Preuzeto s: https://github.com/clarinsi/classla.

Google Translate - Google Search. Preuzeto s: https://www.google.com/search?q=Google+Translate&rlz=1C1GCEA_enHR859HR859&oq=Google+Translate&aqs=chrome..69i57j69i59j0i271l2j69i64.8907j0j4&sourceid=chrome&ie=UTF-8.

HR4EU. Preuzeto s: https://www.hr4eu.hr.

Hrportali. Preuzeto s: https://www.hrportali.com/.

Hrvatska enciklopedija. Preuzeto s: https://enciklopedija.hr/.

Hrvatska jezična riznica, Institut za hrvatski jezik i jezikoslovlje. Preuzeto s: http:// riznica.ihjj.hr/dokumentacija/index.hr.html.

Hrvatski jezični portal. Preuzeto s: https://hjp.znanje.hr/index.php?show=search_by_id&id=f19uXhV0.%C5%BE.

Hrvatski jezik na internetu - JEZIK.HR. Preuzeto s: https://jezik.hr/hrvatski-na-internetu.html.

Hrvatski mrežni rječnik – Pojmovnik. Preuzeto s: http://ihjj.hr/mreznik/page/pojmovnik/6/.

Hrvatski nacionalni korpus. Preuzeto s: https://web.archive.org/web/20160606073223/ http://www.hnk.ffzg.hr/.

hrWaC Corpus Info. Preuzeto s: https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fhrwac22_rft1.

Hunalign – sentence aligner. Preuzeto s: http://mokk.bme.hu/resources/hunalign/.

Interpreta.hr. Preuzeto s: https://www.interpreta.hr/hirek-attekintes/anglizmi-u-hrvatskoj.

Jenset, G. (2008). Basic statistics for corpus linguistics. Uib.No.

Jezikoslovlje | Hrvatska enciklopedija. Preuzeto s: https://enciklopedija.hr/natuknica.aspx?ID=29132.

Kontekst.io. Preuzeto s: https://www.kontekst.io/.

KorpusDK. Preuzeto s: http://ordnet.dk/korpusdk.

London-Lund Corpus of Spoken English (LLC) - Språkbanken. Preuzeto s: https://www.nb.no/sprakbanken/en/resource-catalogue/oai-clarino-uib-no-london-lund/.

Memsource. Preuzeto s: - https://www.google.com/search?q=memsource&rlz=1C1GCEA_enHR859HR859&oq=memsource&aqs=chrome..69i57j0i512j0i20i263i512j0i512l7.1474j0j7&sourceid=chrome&ie=UTF-8.

MULTEXT-East Croatian part-of-speech tagset (version 5). Preuzeto s: https://www.sketchengine.eu/multext-east-croatian-part-of-speech-tagset/.

Nacionalna Sveučilišna knjižnica (NSK). Preuzeto s: https://www.nsk.hr/disertacije-i-magistarski-radovi/.

NoSketchEngine. Preuzeto s: www.clarin.si/noske/run.cgi/corp_info?corpname=hrwac&struct_attr_stats=1.

O hrvatskome jeziku - Institut za hrvatski jezik i jezikoslovlje. Preuzeto s: http://ihjj.hr/stranica/o-hrvatskome-jeziku/26/.

Onion. Corpus Tools. Preuzeto s: https://corpus.tools/wiki/Onion.

Open-source Natural Language Processing tools - Lexical Computing. Preuzeto s: https://www.lexicalcomputing.com/language-databases-tools-solutions/natural-language-processing-tools/.

Regular expressions | Sketch Engine. Preuzeto s: https://www.sketchengine.eu/guide/regular-expressions/.

ReLDI | Regional Linguistic Data Initiative. Preuzeto s: https://reldi.spur.uzh.ch/.

Scott, M. (2008). WordSmith Tools version 5, Liverpool: Lexical Analysis Software.

Sintaktička i semantička analiza dopuna i dodataka u hrvatskom jeziku (SARGADA) - Institut za hrvatski jezik i jezikoslovlje. Preuzeto s: http://ihjj.hr/projekt/sintakticka-i-semanticka-analiza-dopuna-i-dodataka-u-hrvatskom-jeziku-sargada/90/.

Sketch Engine. Preuzeto s: https://www.sketchengine.eu/.

Statistics in Corpus Linguistics: Lancaster Stats Tools Online. Preuzeto s: http://corpora.lancs.ac.uk/stats/toolbox.php.

Stemmer for Croatian | Natural Language Processing group. Preuzeto s: http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/.

Struna | Hrvatsko strukovno nazivlje (ihjj.hr). Preuzeto s: http://struna.ihjj.hr/.

User Guide | Sketch Engine. Preuzeto s: https://www.sketchengine.eu/guide/.

Wmatrix corpus analysis and comparison tool. Preuzeto s: https://ucrel.lancs.ac.uk/wmatrix/.

Word embeddings CLARIN.SI-embed.hr 1.0. Preuzeto s: https://www.clarin.si/repository/xmlui/handle/11356/1205.

KAZALO IMENA

KAZALO POJMOVA

BILJEŠKA O AUTORICI

Mirjana Borucinsky zaposlena je kao izvanredna profesorica na Katedri za strane jezike Pomorskoga fakulteta Sveučilišta u Rijeci, gdje drži kolegije *Engleski jezik* i *Njemački jezik* (struke). Na preddiplomskome studiju anglistike pri Filozofskome fakultetu Sveučilišta u Rijeci od 2015. godine drži kolegij *Uvod u prevođenje.* Njezina su glavna područja znanstvenoga interesa korpusna lingvistika, jezične tehnologije, jezik struke i prevođenje. U koautorstvu je objavila knjigu *Notes on Written Communication in Marine Engineering* (Sveučilište u Rijeci, 2020). Članica je *Centra za jezična istraživanja* (CJI) Sveučilišta u Rijeci, *Hrvatskoga društva za primijenjenu lingvistiku* (HDPL), *Udruge nastavnika jezika struke na visokoškolskim ustanovama* (UNJSVU) te *International Maritime Lecturer's Association* (IMLA). Dobitnica je nagrade Hrvatskoga društva za primijenjenu lingvistiku za najuspješnije izlaganje mladih istraživača u području primijenjene lingvistike na znanstvenome skupu „Metodologija i primjena lingvističkih istraživanja" (Zadar, 2015).

Autorica je trenutno zaposlena u svojstvu istraživača na projektima *Engleske riječi u hrvatskome jeziku: identifikacija, afektivno-semantičko normiranje i ispitivanje kognitivne obrade bihevioralnim i neuroznanstvenim metodama* (HRZZ, 2020 – 2025), UNIRI CLASS A1 *Otvoreno personalizirano obrazovanje - Jezične tehnologije i digitalna obrada teksta*, te u svojstvu voditelja na projektu UNIRI CLASS A2 *Digitalno građanstvo - inovacije u učenju i poučavanju (2022.) - Razvoj komunikacijskih vještina s pomoću jezičnih tehnologija*.

SVEUČILIŠTE U RIJECI
POMORSKI FAKULTET U RIJECI

ISBN 978-953-165-140-0