4.

OBRADA PRIRODNOG JEZIKA, RAČUNALNA PRAGMATIKA I KORPUSNOPRAGMATIČKI PRISTUP JEZIKU U UPOTREBI

Ova cjelina knjige posvećena je računalnoj pragmatici – mladoj disciplini koja se bavi teorijskim i praktičnim aspektima razvoja i primjene računalnih resursa i alata za potrebe proučavanja jezika u upotrebi, a povezuje višestruka znanstvena područja (prvenstveno računarstvo, informacijske i komunikacijske znanosti te lingvistiku). Računalna pragmatika „pokriva" razne istraživačke teme, a svojim rezultatima pridonosi razvoju svih jezgrenih područja pragmatike. U poglavljima koja slijede cilj nam je ponuditi opći pregled ove discipline te detaljnije predstaviti pojedina računalnopragmatička tematska područja koja su povezana s istraživanjima provedenim u nastavku ove knjige (Cjeline 6–9).

Cjelina je organizirana u dva dijela. U prvome se dijelu računalna pragmatika smješta u okvire obrade prirodnog jezika (u nastavku teksta OPJ). U poglavljima koja uključuje predstavljeno je čime se sve bavi OPJ, definirani su temeljni pojmovi poput prirodnih, umjetnih i formalnih jezika te je prikazan odnos između OPJ-a i (računalne) lingvistike. Drugi dio cjeline posvećen je prikazu računalne pragmatike – njezina razvoja, istraživačkih ciljeva i područja bavljenja. U prikazu je posebna pažnja posvećena dvama tematskim područjima računalne pragmatike: jezičnim korpusima te problemu interpretacije i generiranja govornih činova. U poglavlju posvećenome jezičnim korpusima i korpusnopragmatičkome pristupu jeziku u upotrebi ponuđen je opći prikaz korpusa, predstavljena su njihova obilježja, podjela na vrste te mogućnosti njihove primjene. Nadalje, prikazane su pojedine sheme za označavanje korpusa, kao i alati za njihov razvoj i analizu. Osim toga, prikazani su izabrani javno dostupni korpusi hrvatskoga i srpskoga jezika te primjer jednoga korpusa s pragmatičkom anotacijom. Zasebno je poglavlje posvećeno korpusnoj pragmatici – njezinu razvoju te prikazu korpusnopragmatičkoga pristupa jeziku u upotrebi. U završnome poglavlju cjeline predstavljeni su računalnopragmatički pristupi interpretaciji i generiranju govornih činova.

4.1. Obrada prirodnog jezika

4.1.1. Definicija i temeljni pojmovi

Pojam *obrada prirodnog jezika* (OPJ) (eng. *Natural Language Processing*, *NLP*)[59] potječe iz područja računarstva, a odnosi se na interdisciplinarno područje koje se bavi praktičnim aspektima primjene računala za potrebe izvršavanja zadataka koji uključuju ljudski, *prirodni* jezik, i to s dva moguća cilja: (a) omogućavanje i/ili unaprjeđivanje komunikacije između čovjeka i računala ili čovjeka i čovjeka; (b) jednostavna i složena računalna obrada teksta i govora (Jurafsky i Martin 2009: 1).

Prirodni jezik (eng. *natural language*) „svaki [je] jezik kojim se govori i koji je nekome materinski jezik čija su pravila nastala spontano i evolucionistički" (Jojić 2015, natuknica *Jezik*), a služi za međuljudsku komunikaciju (npr. hrvatski, engleski, japanski itd.). Prirodni jezik primarno se ostvaruje u govorenome obliku (na zvučnoj razini), dok se u pisanome obliku ostvaruje na grafičkoj razini.

Osim prirodnih jezika postoje i tzv. umjetni jezici (eng. *constructed* ili *artificial languages*) – razni komunikacijski sustavi koje su ljudi razvili ciljno, za točno određenu svrhu. Takvi jezici, dakle, nisu nastali spontano poput prirodnih jezika. Jedan od najpoznatijih umjetnih jezika jest esperanto. Među umjetne jezike ubrajaju se primjerice i sustavi za vizualni prikaz glazbe te prometni znakovi kojima se ostvaruje jednoznačna komunikacija u prometu. Programski jezici, kao skupovi rezerviranih riječi i simbola kojima se pišu naredbe računalima za izvršavanje određenih računalnih zadataka, također su primjer umjetnih jezika razvijenih radi omogućavanja jednoznačne komunikacije između čovjeka i računala.

U području OPJ-a, jednako kao i u logici, matematici i računarstvu, prirodnim su jezicima suprotstavljeni tzv. formalni jezici (eng. *formal languages*), a to su umjetni jezici koji su formalizirani. Formalizirani su oni jezici čija je sintaksa precizno i algoritamski definirana (Crespi Reghizzi, Breveglieri i Morzenti 2019: 5), a obilježava ih jednoznačnost u komunikaciji: određeni niz znakova ili pripada tom formalnom jeziku ili ne pripada. Primjeri formalnih jezika jesu silogizam, Booleova algebra, programski jezik poput Pythona i sl.

[59] Slično se značenje u engleskome jeziku pripisuje i pojmu *(computer) speech and language processing*. OPJ-u je srodna računalna lingvistika (eng. *computational linguistics*, *CL*), koja se bavi primjenom računala za potrebe lingvističkih istraživanja. Obradu prirodnog jezika i računalnu lingvistiku povezuje pojam *jezične tehnologije* (eng. *(human) language technology*, (*H*)*LT*), koji se odnosi na tehnologije koje se bave računalnom obradom prirodnojezičnih podataka. Više o korpusnoj pragmatici i jezičnim tehnologijama v. u Poglavlju 4.2.

Prirodni i formalni jezici slični su na strukturnome planu (prema Crespi Reghizzi, Breveglieri i Morzenti 2019: 7–8). Podudarno prirodnim jezicima, formalni se jezici sastoje od abecede[60] (tj. skupa znakova), skupa riječi (tj. nizova znakova iz abecede jezika) i skupa pravila za ispravno stvaranje nizova znakova iz abecede jezika. Nadalje, formalni se jezici sastoje od leksikona ili vokabulara (tj. skupa dozvoljenih riječi u jeziku) te od skupa rečenica (tj. skupa pravila za ispravno stvaranje nizova riječi iz leksikona jezika). Međutim, prirodni i formalni jezici razlikuju se u jednoj temeljnoj karakteristici, a to je da su prirodni jezici višeznačni (eng. *ambiguity*), dok su formalni jezici uglavnom razvijani radi ostvarivanja jednoznačnosti u komunikaciji. Upravo je razlučivanje višeznačnosti (eng. *disambiguation*) jedan od središnjih ciljeva OPJ-a (v. Poglavlje 4.1.2.1).

4.1.2. Predmet bavljenja i ciljevi

OPJ je interdisciplinarno područje koje je obuhvaćeno (barem) četirima znanstvenim područjima: tehničkim znanostima (prvenstveno računarstvom), prirodnim znanostima (raznim granama matematike), društvenim znanostima (prvenstveno informacijskim i komunikacijskim znanostima) te humanističkim znanostima (prvenstveno lingvistikom). Predmet proučavanja OPJ-a podaci su zakodirani prirodnim jezikom u formi teksta ili govora. Prirodnojezični podaci nestrukturirani su podaci koje je potrebno oblikovati i strukturirati u informacije i znanje. Jedan od temeljnih izazova OPJ-a formalizacija je prirodnoga jezika za potrebe računalne obrade jezičnih podataka bez gubitka informacija (Bago 2014a: 15).

Aplikacije za obradu prirodnog jezika od ostalih se aplikacija za obradu podataka razlikuju po tome što se zasnivaju na znanju o prirodnome jeziku. Primjerice, aplikacija koja računa broj znakova[61] u tekstu ne zahtijeva nikakvo jezično znanje, dok aplikacija koja računa broj riječi u tekstu treba znati prepoznati riječ. To je primjer vrlo ograničenoga jezičnoga znanja, dok složenije aplikacije zahtijevaju šire i dublje jezično znanje (Jurafsky i Martin, 2009: 2). Aplikacije mogu biti samostalne (kao npr. *Ispravi.me*[62], *online*-provjernik pravopisa hrvatskoga jezika) ili mogu biti sastavnice drugih programa ili aplikacija (npr. provjernik pravopisa hrvatskoga jezika u sklopu MS Office Word aplikacije). Osim provjernicima pravopisa (eng. *spell checker*) OPJ se bavi raznim drugim istraživačkim temama – od optičkoga prepoznavanja znakova (eng. *optical character recognition*, *OCR*), tokenizacije (eng. *tokenization*), lematizacije (eng. *lemmatization*), prepoznavanja govora (eng. *speech*

[60] Pojam *abeceda* ovdje se odnosi na skup svih znakova i simbola koji su mogući u nekome formalnome jeziku.

[61] Ovdje koristimo termin *znak* u računalnome smislu, koji podrazumijeva slova, znamenke, interpunkcijske znakove, bjeline i sl.

[62] https://ispravi.me/

recognition), sustava za diktiranje (eng. *speech-to-text*) i pretvorbe teksta u govor (eng. *text-to-speech*) do automatskoga sažimanja teksta (eng. *automatic text summarization*), dijaloških sustava (eng. *dialogue system*), strojnoga prevođenja (eng. *machine translation*) i izrade sustava uvježbanih za odgovaranje na pitanja (eng. *question answering*).

Iako se sve navedene istraživačke teme OPJ-a neposredno ili posredno mogu povezati s računalnom pragmatikom, u narednim ćemo se poglavljima usmjeriti na teme koje su na teorijsko-metodološkome planu povezane s korpusnopragmatičkim istraživanjima predstavljenima u nastavku knjige.

*

OPJ može biti usmjeren na razumijevanje (eng. *natural language understanding*, *NLU*) i na generiranje prirodnoga jezika (eng. *natural language generation*, *NLG*).

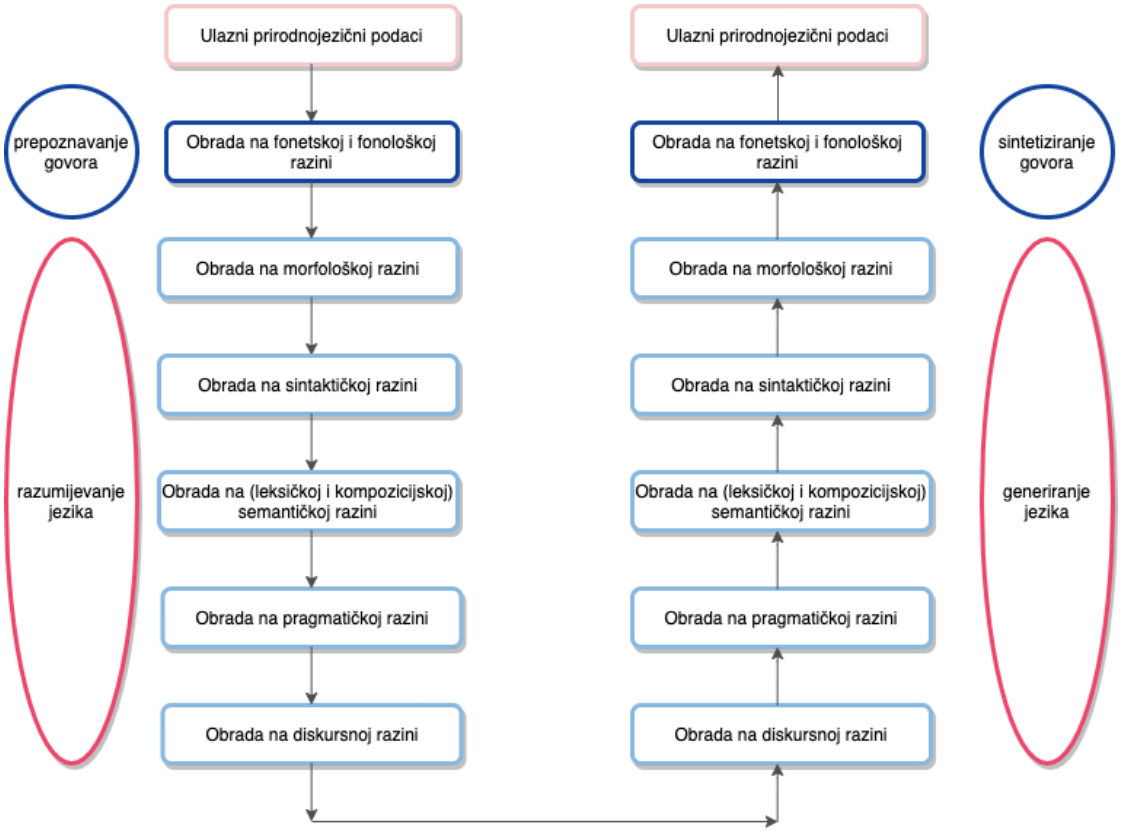
Kao i u tradicionalnoj lingvistici prirodnojezično se znanje u OPJ-u može podijeliti na različite jezične razine: fonetsku, fonološku, morfološku, sintaktičku, semantičku, pragmatičku i diskursnu. Za svaku od navedenih razina jezičnoga znanja mogu se razviti računalni modeli i algoritmi koji se bave razlučivanjem višeznačnosti, a više razine često ovise o rezultatima nižih razina. Tako je, primjerice, nemoguće razviti dijaloški sustav za hrvatski jezik ako ne postoje jezične tehnologije za hrvatski jezik na nižim razinama.

Jurafsky i Martin (2009: 2–4) na primjeru dijaloga iz igranoga filma Stanleyja Kubricka *2001: Odiseja u svemiru* (1968) pojašnjavaju što je sve potrebno da bi računalo HAL 9000 – vjerojatno jedan od najpoznatijih (doduše izmišljenih) dijaloških sustava – moglo komunicirati na engleskome jeziku sa članovima postaje svemirske letjelice *Discovery One*:

> **Dave Bowman:** *Otvori vrata za pristanak kapsule, HAL-e.*
> **HAL:** *Žao mi je, Dave. Bojim se da to ne mogu učiniti.*
> **Dave Bowman:** *U čemu je problem?*
> **HAL:** *Mislim da znaš u čemu je problem, kao i ja.*

Da bi dijaloški sustav poput HAL-a razumio jezik i govorio, treba imati sposobnost obrade ulaznih i izlaznih prirodnojezičnih podataka na različitim jezičnim razinama. Na fonetskoj i fonološkoj razini, uslijed obrade ulaznih prirodnojezičnih podataka i prepoznavanja govora, sustav treba prepoznati govor, odnosno iz zvučnih signala segmentirati govor na manje jedinice kao što su riječi, slogovi i glasovi. Uslijed obrade izlaznih prirodnojezičnih podataka i sintetiziranja govora sustav treba na temelju pisanoga iskaza sintetizirati zvučni val koji zvuči prirodno. Na morfološkoj

razini, uslijed obrade ulaznih prirodnojezičnih podataka, sustav treba prepoznati morfološke kategorije (npr. razlikovati jedninske i množinske oblike, prepoznati padeže i sl.), dok uslijed obrade izlaznih prirodnojezičnih podataka treba generirati ispravne oblike riječi. Na sintaktičkoj razini, uslijed obrade ulaznih prirodnojezičnih podataka, sustav treba prepoznati koje su riječi povezane u višerječne izraze te prepoznati rečenice i klasificirati ih prema tome jesu li izjavne, upitne ili usklične. Na istoj razini, uslijed obrade izlaznih prirodnojezičnih podataka, sustav treba generirati gramatički ispravne rečenice odgovarajućega značenja. Na semantičkoj razini, uslijed obrade ulaznih podataka, sustav treba razumjeti značenje riječi izvan konteksta i u kontekstu (leksička semantika, eng. *lexical semantics*) te značenja njihovih sintagmatskih spojeva (kompozicijska semantika, eng. *compositional semantics*). Na primjer, sustav treba prepoznati da Dave koristi višeznačnu riječ *kapsula* u tehničkome značenju 'dio svemirske letjelice', a ne u farmaceutskome značenju 'način doziranja lijekova u prahu ili u tekućemu stanju'. Kod obrade izlaznih prirodnojezičnih podatka sustav treba odabrati odgovarajuću riječ ili kombinaciju riječi koja će prenijeti planirano značenje. Na pragmatičkoj razini, uslijed obrade ulaznih prirodnojezičnih podataka, tj. razumijevanja jezika, sustav treba prepoznati odnose između (doslovnoga) značenja iskaza i govornikove komunikacijske namjere. Na primjer, sustav treba prepoznati da je izjavna rečenica *Otvori vrata za pristanak kapsule, HAL-e.* ustvari zahtjev sustavu da otvori vrata. Uslijed generiranja jezika sustav pak treba posjedovati znanje o prihvatljivosti tijeka konverzacije i načelima uljudnosti. Na primjer, nakon što je HAL pravilno interpretirao Daveov zahtjev za otvaranjem vrata kapsule, HAL je ponudio odgovor *Žao mi je, Dave. Bojim se da ne mogu to učiniti.* umjesto jednostavnoga odgovora *Ne*. Na razini diskursa sustav treba posjedovati znanje o jezičnim jedinicama većim od jednoga iskaza, odnosno posjedovati znanje o prethodnim iskazima te o izvanjezičnoj stvarnosti i općemu znanju. Kod obrade ulaznih prirodnojezičnih podataka to se može odnositi na problem razrješavanja koreferencijalnosti (eng. *coreference resolution*), tj. identifikacije referenata u sklopu istoga diskursa, te na problem razrješavanja referencijalnosti, tj. identifikacije referenata iz izvanjezične stvarnosti. Kod generiranja jezika sustav pak treba posjedovati izvanjezično znanje da bi donijelo odluku o ispravnome odgovoru ili replici. Iz HAL-ove izjave *Mislim da znaš u čemu je problem, kao i ja.* vidljivo je da sustav pamti prethodne dijaloge. Na Slici 1 prikazana je shema dijaloškoga sustava.



Slika 1. Shema dijaloškoga sustava.

4.1.2.1. Razrješavanje problema višeznačnosti

U prethodnome smo poglavlju istaknuli da je višeznačnost glavno obilježje po kojemu se prirodni jezici razlikuju od formalnih. Prirodnojezični podaci višeznačni su kada postoji više od jedne jezične strukture kojom se mogu opisati (Jurafsky i Martin 2009: 4). Većina istraživačkih tema iz područja OPJ-a upravo se bavi razrješavanjem višeznačnosti na različitim jezičnim razinama: morfološkoj (Primjer 36a), sintaktičkoj, (Primjer 36b), leksičkoj (Primjer 36c), diskursnoj (Primjer 36d) i pragmatičkoj (Primjeri 36e).

Primjer 36

(a) Došao je bez unuka.
(b) Stariji muškarci i žene skloni su anemiji.
(c) Miš je na stolu.
(d) Ivan ima psa. Stalno maše repom.
(e) Jesi li ponio aspirin?

U Primjeru 36a iskaz se zbog višeznačnosti oblika imenice *unuka* može interpretirati na dva načina: *Došao je bez jednoga unuka / Došao je bez više unuka.* Iskaz iz Primjera 36b zbog sintaktičke se višeznačnosti subjektnoga skupa može interpretirati kao tvrdnja koja se odnosi na starije muškarce i starije žene ili na starije muškarce i sve žene. Imenica *miš* u Primjeru 36c može se odnositi na životinju ili na vanjsku računalnu komponentu. U primjeru 36d uzrok višeznačnosti dvosmislena je koreferencijalnost zbog koje se drugi dio iskaza može protumačiti na dva načina: *Ivan stalno maše repom* i *Pas stalno maše repom.* Primjer 36e može se interpretirati kao upitni iskaz ili kao implicitni zahtjev ili molba: *Daj mi aspirin.*

Shodno Tadiću (2003) jedan od glavnih problema vezanih uz višeznačnost u hrvatskome jeziku predstavlja homografija (istopisnost), a autor je dijeli na unutarnju i vanjsku. Unutarnja homografija javlja se u slučaju morfološke višeznačnosti, a vanjska u slučaju leksičke višeznačnosti:

> Prvu bismo mogli nazvati *unutarnjom istopisnošću* u slučaju kad ista pojavnica (ili niz pismena) predstavljaju različite oblike iste leme tj. oblike s različitim morfosintaktičkim opisima, no još uvijek pripadaju istoj lemi. Primjer za to su dativ, lokativ i instrumental množine u hrvatskome koji najčešće imaju identičan lik, ali predstavljaju zapravo tri različita oblika iste leme s tri različita morfosintaktička opisa [...] Druga vrsta istopisnosti mogla bi se nazvati *vanjskom istopisnošću* kad ista pojavnica (ili niz pismena) predstavlja oblike više različitih lema. (Tadić 2003: 126)

Kao primjer unutarnje homografije Tadić (2003: 126) navodi oblik *gledateljima*, koji je jednak u dativu, lokativu i instrumentalu množine leme *gledatelj*. Kao primjer vanjske homografije autor navodi oblik *cijene*, koji se može odnositi na imenicu *cijena* te je u tome slučaju unutarnje homografan (Gsg=Npl=Apl=Vpl), ali i na 3. lice množine prezenta glagola *cijeniti*. Tadić (ibid. 127) navodi zanimljiv rezultat analize homografa iz flektivnoga leksikona *Hrvatskoga morfološkoga leksikona* koja je pokazala da se vanjska homografija pojavljuje u 1,73 % lema, a unutarnja homografija u 56,21 % lema.

*

U području OPJ-a tri su temeljna pristupa razvoju modela i algoritama za rješavanje problema višeznačnosti na svim jezičnim razinama, a to su: pristup temeljen na pravilima (eng. *rule-based approach*), statistički pristup (eng. *statistical approach*) i duboko učenje (engl. *deep learning*) (prema Jurafsky i Martin 2009: 5). U praksi se često koristi hibridni pristup koji kombinira više pristupa za rješavanje određenoga prirodnojezičnoga problema.

Najstariji je pristup temeljen na pravilima, a prvi je put primijenjen 1950-ih godina za razvoj sustava strojnoga prevođenja s ruskoga na engleski jezik, koji se sastojao od šest gramatičkih pravila i rječnika od 250 riječi (Hutchins 1995). Pod ovim se pristupom podrazumijevaju algoritmi koji se temelje na ručno sastavljenim popisima pravila i riječi. Ovaj pristup najčešće se primjenjuje kada za određeni jezik nisu dostupni jezični resursi kao izvor prirodnojezičnih podataka. Ovaj se pristup također koristi u fazi koja prethodi obradi podataka ili koja slijedi nakon obrade podataka. Glavna prednost sustava temeljenih na pravilima njihova je visoka točnost u slučajevima kada su pravila dobro osmišljena. Međutim, oni imaju i svoje nedostatke: proces sastavljanja pravila zahtjevan je i dugotrajan, ovisni su o pojedinim jezičnim stručnjacima te ne pružaju mogućnost proširivanja jer ono zahtijeva povećanje kompleksnosti pravila.

S vremenom su računala postajala brža, a memorija jeftinija, što je omogućilo prikupljanje sve veće količine prirodnojezičnih podataka i njihovu bržu obradu. Time su kasnih 1980-ih i početkom 1990-ih godina u području OPJ-a postavljeni temelji za razvoj statističkih pristupa koji se oslanjaju na prirodnojezične podatke (Johnson 2009). U sklopu ovoga pristupa najčešće su se primjenjivali algoritmi strojnoga učenja (eng. *machine learning*), koji se automatski unaprjeđuju na temelju podataka, a čiji je cilj razvoj učinkovitih i točnih algoritama za predviđanje (Mohri, Rostamizadeh i Talwalkar 2018: 1). Glavne prednosti sustava temeljenih na statističkome pristupu u tome su što ovise o prirodnojezičnim podacima, brzo uče na temelju iskustva i lako su proširivi novim podacima. Budući da ovi sustavi ovise o kvaliteti ulaznih prirodnojezičnih podataka, njihova je točnost često niža u odnosu na sustave temeljene na pravilima, što može biti njihov nedostatak.

Daljnjim razvojem računala, početkom ovoga tisućljeća, razvija se pristup dubokoga učenja, koji se temelji na (umjetnim) neuronskim mrežama (eng. [*artificial*] *neural networks*, *ANN*). To je pristup koji se koristi u različitim područjima, a intenzivno se počeo primjenjivati u OPJ-u sredinom 2010-ih godina (Goldberg 2017: 4–5). Glavna novost koju donosi ovaj sustav dobri su rezultati na nelinearnim procesima, koji koriste više slojeva u mreži, pri čemu svaki sloj ulazne podatke pretvara u apstraktniji i složeniji prikaz podataka, te ne služi samo za predviđanje već i za postizanje ispravnoga prikaza podataka (ibid. 2). Ovaj pristup također se temelji na prirodnojezičnim podacima, a njegova je glavna prednost samostalnost u učenju – za razliku od statističkoga pristupa, koji zahtijeva značajnu ljudsku intervenciju za identifikaciju i odabir reprezentativnih značajki. Glavni nedostatak ovoga pristupa nepostojanje je teorije koja podržava njegove metode, stoga se potvrde odvijaju isključivo empirijski.

Prema Jurafskyju i Martinu (2009: 5) najbitniji formalni modeli jezičnoga znanja koji se temelje na teoriji, a koji se primjenjuju u obradi prirodnog jezika, jesu: (1) automati stanja (eng. *state machines*) – formalni modeli koji se sastoje od stanja, skupa početnih stanja, ulazne abecede te prijelaznih funkcija između tih stanja (Black[63], natuknica *State machine*); (2) sustavi temeljeni na skupu pravila (eng. *rule-based systems*), čija su pravila iskazana u obliku „ako-onda" naredbi, skupa činjenica i interpretera pravila prema ulaznim činjenicama (Grosan i Abraham 2011: 149); (3) logički sustavi – formalni modeli jezičnoga znanja potekli iz područja logike; (4) vjerojatnosni modeli (eng. *probabilistic models*) – visokoprimjenjivi modeli koji na temelju prirodnojezičnoga uzorka donose pretpostavke o jeziku, a mogu se primijeniti kao proširenje prethodno navedenih triju modela (Jurafsky i Martin 2009: 5); (5) modeli vektorskoga prostora (eng. *vector-space models*), koji reprezentiraju tekst vektorom, pri čemu elementi vektora označavaju pojavljivanje riječi ili izraza u tekstu, a svaka riječ ili izraz u tekstu ima dimenziju (Miner 2012: 45).

4.1.3. Obrada prirodnog jezika i lingvistika

Zahvaljujući sve većoj dostupnosti resursa i alata iz područja OPJ-a njima se u novije vrijeme sve više služe lingvisti (Kurdi 2016: x–xi). Njihova je primjena u lingvistici omogućila empirijsko testiranje postojećih lingvističkih teorija te „otvorila" nove dimenzije istraživanja zasnovanih na ovjerenoj jezičnoj građi (posebice kvantitativnoga tipa). Prema Leechevim riječima:

> Računalo je, kao iznimno snažno tehnološko sredstvo, omogućilo tu novu vrstu lingvistike. Tako je tehnologija (kao što je to već stoljećima u prirodnim znanostima) dobila važniju ulogu od puke podrške i olakšavanja istraživanja. Vidim je kao esencijalno sredstvo za novu vrstu znanja i kao 'Sezame, otvori se' novom načinu razmišljanja o jeziku. (Leech 1992a: 106 prema Tadić 1996: 604)

Lingvistička disciplina koja se bavi primjenom računala za potrebe provođenja lingvističkih istraživanja zove se računalna lingvistika. Pitanje odnosa između računalne lingvistike i OPJ-a podložno je različitim interpretacijama. Dok neki stručnjaci OPJ i računalnu lingvistiku smatraju sinonimnim pojmovima, drugi ih smatraju zasebnim područjima, unatoč tome što se u mnogočemu međusobno preklapaju te se ne mogu jasno razgraničiti. Kao glavna razlika među njima obično se ističe usmjerenost OPJ-a na praktične aspekte razvoja jezičnih tehnologija, dok se računalna lingvistika u većoj mjeri bavi teorijskim aspektima toga zadatka te je usmjerena na pitanja njihove primjene u lingvistici (usp. Kurdi 2016; Stabler 2003). Takvo je tumačenje u skladu s opisom dvostruke istraživačke motivacije računalnih

[63] https://xlinux.nist.gov/dads/

lingvista, prikazane u sklopu definicije računalne lingvistike koju nudi organizacija *Associatian for Computational Linguistics*:

> Računalna lingvistika bavi se znanstvenim proučavanjem jezika iz računalne perspektive. Računalni lingvisti bave se izradom računalnih modela različitih jezičnih pojavnosti. […] Računalnolingvistički napori mogu biti znanstveno ili tehnološki motivirani, tj. mogu biti usmjereni na potragu za računalnim objašnjenjima pojedinih lingvističkih fenomena ili na samu izradu radnih komponenti prirodnojezičnih sustava. (*Associatian for Computational Linguistics*)[64]

Jezične tehnologije predstavljaju krovni pojam koji povezuje OPJ i računalnu lingvistiku. Jezične tehnologije, kao tehnologije koje se bave računalnom obradom prirodnojezičnih podataka na svim jezičnim razinama, dijele se na jezične resurse i jezične alate[65] (prema Tadić 2003: 27).

> Jezični su resursi računalno pribavljene, pohranjene i podržane zbirke jezičnih podataka, a sastoje se ponajprije od korpusa, a potom od rječnika. Jezični resursi služe ili za razvitak novih jezičnih resursa (npr. specijaliziranog potkorpusa iz nekog općeg korpusa, rječnika na temelju korpusa itd.), ili za razvitak novih alata (npr. sustavi za segmentaciju na rečenice temeljeni na evidenciji iz korpusa ili na leksikonima/popisima riječi koje obvezatno počinju velikim slovom itd.). (ibid. 28)

U narednim poglavljima najviše ćemo se usredotočiti na korpuse. Budući da korpusi u drugoj polovici 20. stoljeća postaju glavni računalni jezični resurs kojima se služe lingvisti, razvija se korpusna lingvistika kao zasebna lingvistička disciplina. Budući da se OPJ i računalna lingvistika bave razvojem i izradom korpusa, korpusna je lingvistika usko povezana s njima – ali i s brojnim drugim lingvističkim disciplinama:

> Za razliku od ostalih jezikoslovnih disciplina, korpusna lingvistika određena je ne toliko područjem istraživanja koliko metodološkom osnovicom na kojoj se temelji istraživanje. Stoga se korpusni pristup (ili korpusna metodologija) lako može primijeniti u različitim lingvističkim disciplinama: fonologiji, morfologiji, sintaksi, sociolingvistici, kognitivnoj lingvistici itd., i to najčešće u kombinaciji s drugim, tim disciplinama inherentnim, metodološkim postupcima. (Tadić 1996: 604)

Izradom korpusa te drugih alata i resursa za potrebe pragmatičkih istraživanja bavi se računalna pragmatika, dok se korpusnim pristupom pragmatičkim fenomenima bavi korpusna pragmatika. U narednim poglavljima slijedi njihov pobliži prikaz.

---

[64] V. poveznicu https://www.aclweb.org/portal/.

[65] Tadić navodi i treću sastavnicu jezičnih tehnologija, a to su komercijalni proizvodi. Takvu podjelu smatramo nespretnom s obzirom na to da jezični resursi i jezični alati mogu biti komercijalni proizvodi.

4.2. Računalna pragmatika

4.2.1. Razvoj i predmet bavljenja

Prema Buntovoj (2020) definiciji računalna je pragmatika vrsta pragmatike koja koristi računalo kao alat u svojim istraživanjima. Riječ je o mladome interdisciplinarnome području koje se bavi izradom modela procesa upravljanja dijalogom i zbirki podataka o uporabi jezika, shemama i standardima za označavanje jezičnih resursa, programskim alatima za izgradnju, označavanje i proučavanje korpusa te modelima procesa generiranja i interpretiranja jezika. Za potrebe izrade modela procesa generiranja i interpretiranja kontekstno ovisnih iskaza bavi se reprezentacijom konteksta i metodama zaključivanja. Računalna pragmatika proučava jezik u pisanome i govorenome obliku, u monološkoj i dijaloškoj formi. Proučavanje jezika u dijaloškoj formi iznimno je izazovan zadatak jer se tijekom konverzacije informacije vezane za iskaze govornika neprestano mijenjaju. Te promjene nastupaju uslijed razmjene informacija među sudionicima komunikacije, stoga se kontekst dijaloga kontinuirano ažurira.

Bunt (ibid.) začetke računalne pragmatike smješta na sam kraj 20. stoljeća, kada Allen (1995) po prvi puta u udžbenik o računalnoj lingvistici uključuje poglavlja posvećena računalnoj pragmatici. U posljednja dva desetljeća ova se disciplina ubrzano razvija. Ipak, još uvijek ne postoje udžbenici i/ili monografije posvećene njezinu općemu prikazu. Dostupne su, međutim, studije primarno posvećene pojedinim računalnopragmatičkim temama, koje između ostaloga uključuju i opće prikaze discipline, stoga predstavljaju svojevrsne „uvodnike” u računalnu pragmatiku (npr. Bunt i Black 2000; Jurafsky i Martin 2009).

Računalna pragmatika bavi se sustavnim proučavanjem relacija između iskaza i konteksta njihove upotrebe pristupajući im iz računalne perspektive (Huang 2017: 6). Računalnopragmatička istraživanja usmjerena su na utvrđivanje odnosa između: (1) iskaza i čina djelovanja; (2) iskaza i ostatka diskursa te (3) iskaza i okolnosti u kojima je iskaz upotrijebljen (ibid.). Shodno Huangu računalna pragmatika ovim odnosima pristupa iz dviju perspektiva:

> S jedne strane, s obzirom na jezični izraz, računalna pragmatika nastoji odrediti na koji se način mogu računalno obraditi relevantna obilježja konteksta u kojemu je upotrijebljen iskaz. S druge strane, u slučaju generiranja jezika, zadatak računalne pragmatike jest konstruirati jezični izraz u kojemu su kodirane relevantne kontekstualne informacije te utvrditi kako se na temelju relevantnih kontekstualnih informacija mogu računalno obraditi relevantna svojstva iskaza. Glavni cilj računalne pragmatike uspostavljanje je eksplicitne računalne reprezentacije ovih odnosa. (ibid.)

Računalna pragmatika doprinosi unapređenju opće pragmatike razvojem računalnih modela interpretacije, generiranja, zaključivanja i učenja, a posebice stvaranjem računalnih alata i resursa za potrebe provođenja pragmatičkih istraživanja. Posebno važno mjesto među jezičnim resursima imaju korpusi s pragmatičkom anotacijom, koji omogućavaju provođenje korpusnopragmatičkih analiza. Osim toga, takvi se korpusi primjenjuju pri izradi sustava za obradu prirodnog jezika primjenom tehnika strojnoga učenja (Bunt 2017: 344).

Računalna pragmatika bavi se svim pragmatičkim jezgrenim područjima, s time da su dosadašnja računalnopragmatička istraživanja u najvećoj mjeri usmjerena na proučavanje implicitnih značenja, govornih činova i analizu konverzacije. Bunt (ibid.) izdvaja tri tematska područja kojima se danas sustavno bavi računalna pragmatika: (1) utvrđivanje mehanizama izvođenja zaključaka te uspostavljanja odgovarajućih interpretacija smisla iskaza; (2) modeliranje jezične upotrebe kao vida djelovanja; (3) utvrđivanje odnosa u sklopu dijaloga/diskursa.

Računalnopragmatička istraživanja koja se bave interpretacijom smisla iskaza zasnivaju se na pretpostavci da se izvođenje zaključaka u jezičnoj komunikaciji primarno zasniva na abdukciji – vrsti logičkoga zaključivanja, tj. deriviranja hipoteza koje potencijalno objašnjavaju neko opažanje:

> Primjerice, iz opažanja da je ulica mokra te da se ulica smoči kada pada kiša proizlazi hipoteza da pada kiša. Hipoteze derivirane abdukcijom nisu logički valjane te mogu biti netočne. Primjerice, ulica može biti mokra jer je pukla vodovodna cijev […] Abdukcija je vrsta logičkoga zaključivanja koje ljudi stalno primjenjuju ne bi li interpretirali i objasnili ono što vide i čuju. (Bunt 2017: 329)

Abdukcija je vrsta silogizma „kojemu je druga premisa samo vjerojatna pa je i zaključak takav" (Klaić 2007, natuknica *Abdukcija*), a koji se 1990-ih godina razvojem računala počeo sve više proučavati i primjenjivati za potrebe programiranja u raznim područjima, pa tako i u OPJ-u (Flach i Kakas 2000: ix, Bylander et al. 1991, Fox 1992). Ova vrsta logičkoga zaključivanja u računalnopragmatičkim se teorijama koristi kao polazište za tumačenje raznih kontekstualno uvjetovanih pragmatičkih fenomena, a posebice mehanizama de/kodiranja konverzacijskih implikatura. Dosadašnja istraživanja iz ovoga područja rezultirala su izradom alata i resursa koji omogućavaju nove uvide u mehanizme prijenosa implicitnih značenja u jezičnoj upotrebi.

Računalnopragmatička istraživanja koja se bave modeliranjem jezične upotrebe kao vida djelovanja temelje se na teoriji dijaloških činova (Bunt 1969, 1989, 2017). Prema Buntovoj definiciji (2017: 332) dijaloški je čin sastavnica komunikacijske aktivnosti sudionika konverzacije koja ima određenu komunikacijsku funkciju i

semantički sadržaj. Teorija dijaloških činova razlikuje se od Austinove i Searleove teorije govornih činova po sljedećim stavkama: (a) dok se teorija govornih činova bavi verbalnim ponašanjem, teorija dijaloških činova uzima u obzir i neverbalno te multimodalno ponašanje; (b) prema teoriji govornih činova jedan iskaz kodira jedan govorni čin, dok su iskazi prema teoriji dijaloškoga čina polifunkcionalni; (c) prema teoriji dijaloških činova tijekom konverzacije informacijska stanja njezinih sudionika neprestano se ažuriraju; (d) dijaloški činovi neodvojivi su od drugih dijaloških činova s kojima su povezani na semantičkoj ili pragmatičkoj razini (ibid.).

Računalnopragmatička istraživanja iz ovoga tematskoga područja usmjerena su na podjelu dijaloga na funkcionalne segmente, koji su po svojoj definiciji minimalne sastavnice komunikacijskoga ponašanja s određenom komunikacijskom funkcijom ili više njih (Geertzen et al. 2007). Osim na segmentiranje dijalogā istraživanja su usmjerena i na određivanje komunikacijskih funkcija njihovih funkcionalnih segmenata, kao i na interpretaciju njihova smisla.

Računalnopragmatička istraživanja koja se bave utvrđivanjem odnosa u sklopu dijaloga/diskursa usmjerena su na uvjetovanost smisla dijaloških činova raznim semantičkim, funkcionalnim i retoričkim čimbenicima.

U okviru ove knjige nije moguće ponuditi detaljan prikaz svih tematskih područja kojima se bavi računalna pragmatika, stoga u nastavku slijede pobliži prikazi dvaju tematskih područja koja su najuže vezana za istraživanja provedena u narednim cjelinama knjige: (1) računalna pragmatika i korpusi; (2) računalnopragmatički pristupi interpretaciji i generiranju govornih činova.

4.2.2. Računalna pragmatika i korpusi

Ovo poglavlje sastoji se od dva dijela: (1) U prvome, općemu dijelu predstavljena je definicija korpusā te su prikazana njihova temeljna obilježja i mogućnosti primjene. Osim toga, ponuđena je njihova podjela na vrste s obzirom na različite kriterije. Posebna su potpoglavlja posvećena označavanju korpusā te alatima za njihov razvoj i analizu. Na koncu su predstavljeni izabrani korpusi hrvatskoga i srpskoga jezika te jedan primjer korpusa s pragmatičkom anotacijom. (2) U drugome dijelu poglavlja prikazan je razvoj korpusne pragmatike te su predstavljena obilježja korpusnopragmatičkoga pristupa jeziku u upotrebi.

4.2.2.1. Definicija, obilježja i primjena korpusa

Korpus (eng. *corpus*) je zbirka jezičnih odsječaka u elektroničkome obliku, odabranih prema vanjskim kriterijima s ciljem prikaza (reprezentacije) jezika ili jezičnih varijanti kao izvora za lingvistička istraživanja (Sinclair 2004).

Korpusi su izvori empirijskih prirodnojezičnih podataka. Oni omogućavaju donošenje zaključaka na temelju prirodnojezičnih podataka, a ne intuicije ili sekundarnih izvora (Svensén 2009: 45). Također omogućavaju kvantifikaciju prirodnojezičnih fenomena: primjerice, primjenom korpusa moguće je izračunati frekvencijske distribucije pojavnica ili izraza, ali i vrsta riječi ako korpus sadržava takve oznake. Korpusi omogućavaju proučavanje tipičnih prirodnojezičnih fenomena i njihovih međuodnosa. Oni, međutim, nisu dobar izvor za proučavanje rijetkih jezičnih fenomena (ibid.). Prema Svensénu (ibid. 51–52) jedan od temeljnih problema do kojega može doći pri korpusnoj analizi izostanak je potvrde nekoga prirodnojezičnoga fenomena za koji znamo da postoji. Taj je problem obično rješiv povećanjem korpusa ili primjenom nekih drugih izvora prirodnojezičnih podataka. U svakom slučaju, valja imati na umu da se korpusnom analizom u pravilu ne može potvrditi nepostojanje određenoga prirodnojezičnoga fenomena.

Prema Sinclairu (2004) korpusi bi trebali zadovoljavati 10 osnovnih načela – načelo autentičnosti, reprezentativnosti, orijentacije, kriterija odabira, metapodataka, uzorkovanja, dokumentacije, uravnoteženosti, teme i homogenosti. (1) Prema načelu autentičnosti sadržaj korpusa treba odražavati komunikacijske obrasce jezične zajednice, stoga treba biti odabran prema njegovoj komunikacijskoj funkciji, a ne prema jezičnome sadržaju. Ovo načelo osigurava da je korpus sastavljen prema vanjskim kriterijima ostvarivanja jezične upotrebe (npr. s obzirom na prigode i situacije u kojima se ostvaruje komunikacija i njezine sudionike), a ne prema unutarnjim kriterijima koji se tiču jezičnih uzoraka unutar jezičnih odsječaka. (2) Načelo reprezentativnosti zasniva se na pretpostavci da korpusi predstavljaju uzorak prirodnojezičnih podataka koji je reprezentativan u smislu da odražava obilježja proučavanoga jezika. (3) Načelo orijentacije odnosi se na mogućnost promatranja i uspoređivanja neovisnih sastavnica korpusa. Zahvaljujući ovome načelu procjena prikladnosti korpusa za planirana prirodnojezična istraživanja leži na samome istraživaču. (4) Prema načelu kriterija odabira prilikom utvrđivanja strukture korpusa potrebno je odrediti nepreklapajuće kriterije odabira jezičnih odsječaka koji će ući u korpus (npr. broj jezika, pisani ili govoreni oblik jezika, stil, razdoblje i sl.). (5) Načelo metapodataka odnosi se na eksplicitno razlikovanje sadržaja korpusa od njegovih opisnih, strukturnih, administrativnih i drugih podataka (npr. izvori jezičnih isječaka, prijelom dokumenata, anotacija i sl.). (6) Načelo uzorkovanja nalaže da treba težiti tome da jezični odsječci sadržavaju cijele dokumente ili transkripte kada je to

moguće. Shodno ovome načelu sastavljači korpusa unaprijed trebaju odrediti minimalnu veličinu korpusa potrebnu za proučavanje određenih prirodnojezičnih fenomena. Načelo uzorkovanja usko je povezano s načelom reprezentativnosti i uravnoteženosti. (7) Načelo dokumentacije odnosi se na vođenje iscrpne dokumentacije o korpusu koja sadržava informacije o dizajnu, sastavu i sadržaju korpusa, kao i obrazloženja o donesenim odlukama prilikom izrade korpusa. Ovo načelo istraživačima omogućava donošenje ispravnih odluka o prikladnosti korpusa za potrebe njihovih istraživanja, kao i ispravno interpretiranje rezultata i zaključivanje na temelju korpusnih istraživanja. (8) Načelo uravnoteženosti nalaže da omjer odabranih sastavnica korpusa treba biti što uravnoteženiji kako jedna sastavnica ne bi imala neprimjeren utjecaj na rezultate upita (npr. dugi tekstovi u malim korpusima imaju velik utjecaj na rezultate upita). (9) Načelo teme nalaže da jedino vanjski kriteriji, a ne unutarnji, trebaju vrijediti prilikom odabira tema sadržaja korpusnih sastavnica. (10) Načelo homogenosti odnosi se na kriterij uključivanja tekstova u korpus, uslijed kojega sastavljači korpusa istodobno trebaju voditi računa o dobroj „pokrivenosti" te o izbjegavanju nepodobnih tekstova. Ovo načelo nalaže sastavljačima korpusa da se prilikom izbora tekstova koji će ući u korpus ne oslanjaju slijepo samo na objektivne kriterije već da primjene i svoje stručno znanje i zdrav razum.

Korpusna se istraživanja dijele na dva temeljna pristupa: korpusno utemeljen pristup (eng. *corpus-based approach*) i korpusom vođen pristup (eng. *corpus-driven approach*). Tognini-Bonelli (2001: 10–11) navodi da korpusno utemeljen pristup podrazumijeva primjenu korpusa kao repozitorija primjera koji služe za razjašnjenje, testiranje ili oprimjerivanje određenih teorijskih tvrdnji (*teorija → podaci*), dok se kod primjene korpusom vođenoga pristupa teorijske tvrdnje oblikuju isključivo na temelju potvrda u korpusima (*podaci → teorija*).

Iako prema Sinclairovoj definiciji korpusi služe za lingvistička istraživanja, oni se koriste i u razne druge svrhe. Već smo spomenuli da korpusi služe za razvoj jezičnih tehnologija. Korpusi se također koriste u nastavi jezika (v. npr. Chambers 2010). Prva gramatika temeljena na čak trima korpusima *A Comprehensive Grammar of the English Language* (Quirk et al. 1985: 33) objavljena je 1985. godine. Dvije godine kasnije obavljen je prvi jednojezični rječnik temeljen na korpusu *Collins Cobuild Dictionary of the English Language* (1987) (McCarthy i O'Keeffe 2010). Prvi dvojezični rječnik temeljen na korpusu jest *Oxford-Hachette French Dictionary* objavljen 1994. godine (Svensén 2009: 46). Korpusi se primjenjuju i za potrebe analize medijskoga (v. npr. Jacobi, Van Atteveldt i Welbers 2016) i političkoga diskursa (v. npr. Frantzi, Georgalidou i Giakoumakis 2019) te u raznim istraživanjima u području rodnih studija (v. npr. Zhao et al. 2017).

4.2.2.2. Vrste korpusa

Postoje različite vrste korpusa, a mogu se klasificirati prema raznim kriterijima. U nastavku slijedi popis i kratak opis najosnovnijih vrsta korpusa (prema Bonelli 2010; Lee 2010; web-stranica *Corpus types*[66]).

Prema jezičnoj pokrivenosti korpusi se dijele na opće (eng. *general corpus*) i specijalizirane (eng. *specialized corpus*). Opći korpusi predstavljaju cjelokupan jezik u govorenome i pisanome obliku pokrivajući širok raspon različitih izvora koji reprezentiraju jezik neke jezične zajednice u cjelini. Primjeri hrvatskih općih korpusa: hrWaC (Ljubešić i Klubička 2014), *Hrvatski nacionalni korpus* (Tadić 1996) i *Hrvatski jezični korpus Riznica* (Ćavar i Brozović Rončević 2012). Primjeri korpusa drugih jezika: srpski jezik: srWaC (Ljubešić i Klubička 2014); bosanski jezik: bsWaC (ibid.); slovenski jezik: *Gigafida* (Krek et al. 2020); češki jezik: *Český národní korpus* (Čermak 1997); britanski engleski jezik: *British National Corpus* (Leech 1992b); američki engleski jezik: *Brown Corpus* (Francis i Kučera 1979); njemački jezik: *Deutsches Referenzkorpus* (Lüngen 2017).

Specijalizirani korpusi ograničeni su prema određenome kriteriju te predstavljaju samo jednu jezičnu varijantu. Takvi korpusi služe za istraživanje stručne terminologije, dijalekata, nestandardnih jezičnih varijeteta i sl. Primjeri ove vrste korpusa za hrvatski jezik: korpus hrvatskih objava s Twittera Tweet-hr (Ljubešić et al. 2019) i korpus novinskih portala ENGRI (Bogunović et al. 2021). Neki primjeri specijaliziranih korpusa drugih jezika: korpus srpskih objava s Twittera Tweet-sr (Ljubešić et al. 2017); korpus akademskoga slovenskoga jezika KAS (Erjavec et al. 2019); korpus akademskoga govorenoga engleskoga jezika *The Michigan Corpus of Academic Spoken English* (Simpson et al. 2002), korpus dramskih tekstova na engleskome jeziku *Leuven Drama Corpus* (Geens, Engels i Martin 1975 prema Bonelli 2010).

Prema broju jezika korpusi se dijele na jednojezične (eng. *monolingual corpus*) i višejezične (eng. *multilingual corpus*). Jednojezični korpusi najzastupljenija su vrsta korpusa. Svi prethodno pobrojani korpusi jednojezični su, što znači da sadržavaju tekstove na samo jednome jeziku. Višejezični korpusi sadržavaju korpuse dvaju ili više jezika, a primjenjuju se za potrebe poredbenih lingvističkih istraživanja. Odluke koje se donose prilikom sastavljanja i dizajniranja višejezičnih korpusa jednake su za sve jezike. Višejezični korpusi nadalje se dijele na paralelne (eng. *parallel corpus*) i usporedive (eng. *comparable corpus*). Paralelni korpusi sastoje se od dva ili više jednojezičnih korpusa koji sadržavaju prijevode istoga teksta. Takvi su korpusi

[66] Opis vrsta korpusa koji se mogu pronaći u alatu za razvoj i analizu korpusa *Sketch Engine*: https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

najčešće sravnjeni na razini rečenice, no mogu biti sravnjeni i na razini odlomka. Primjeri paralelnih korpusa koji sadržavaju hrvatski jezik jesu: DGT-TM (Steinberger et al. 2014), koji je sravnjen na razini rečenice; EUR-Lex Corpus (Baisa et al. 2016), koji je sravnjen na razini odlomka te OPUS2 (Tiedemann 2012), koji je sravnjen na razini rečenice. Usporedivi korpusi sastoje se od dvaju ili više jednojezičnih korpusa ili jezičnih varijanti istoga jezika, a sadržavaju slične tekstove. Budući da se ne sastoje od prijevoda istoga teksta, ovi korpusi nisu sravnjeni. Usporedivi korpusi koji sadržavaju hrvatski jezik jesu: CHILDES (Hržica, Kuvač Kraljević i Šnajder 2013) i ParlaMint (Erjavec et al. 2021).

Prema obliku jezičnoga ostvaraja korpusi se dijele na korpuse pisanoga (eng. *written corpus*) i govorenoga (eng. *spoken corpus*) jezika. Korpusi pisanoga jezika sastoje se od pisanih tekstova, a zbog jednostavnijega razvoja mnogo su češći od korpusa govorenoga jezika. S druge strane, korpusi govorenoga jezika sadržavaju transkripte monologa ili konverzacije. Korpusi ove vrste rijetki su zbog zahtjevnosti njihova razvoja. Primjeri korpusa govorenoga jezika: za hrvatski jezik: CHILDES i HrAL (Kuvač Kraljević i Hržica 2016); za engleski jezik: *The Michigan Corpus of Academic Spoken English*.

S obzirom na vremenski razmak između najstarijega i najnovijega teksta koji sadržavaju, korpusi se dijele na dijakronijske (eng. *diachronic corpus*) i sinkronijske (eng. *synchronic corpus*). Dijakronijski korpusi sadržavaju tekstove iz različitih vremenskih razdoblja, stoga omogućavaju proučavanje jezičnih promjena i jezičnoga razvoja kroz vrijeme. Dijakronijski korpusi obavezno uključuju metapodatke o vremenu nastanka teksta, koji omogućavaju pretraživanje prema tome kriteriju. Prvi korpus ove vrste razvijen je za engleski jezik: *The Helsinki Corpus of English Texts* (Kytö i Rissanen 1992), a pokriva razdoblje od 700. do 1700. godine. Dijakronijski korpus hrvatskoga jezika za sada ne postoji. Primjeri dijakronijskih korpusa drugih jezika: *The Corpus of Historical American English* (Davies, Hegedűs i Fodor 2012), koji pokriva razdoblje od 1810. do 2009. godine; *Timestamped JSI Web Corpus* (Bušta et al. 2017) – skup dijakronijskih mrežnih korpusa koji uključuju 18 jezika, a pokrivaju kraće vremensko razdoblje (2014.–2020.). Sinkronijski korpusi sadržavaju tekstove iz istoga vremenskoga razdoblja. Mrežni korpusi često pripadaju ovoj vrsti korpusa jer se njihov sadržaj prikuplja unutar nekoliko mjeseci. Primjeri takvih korpusa jesu hrWaC te skup TenTen korpusa (Jakubíček et al. 2013) razvijenih za 41 jezik.

S obzirom na kriterij (ne)promjenjivosti korpusi se dijele na statične (eng. *static corpus*) i monitor-korpuse (eng. *monitor corpus*). Statični korpusi su korpusi čiji se sadržaj ne mijenja i čije je sastavljanje završeno. Većina korpusa pripada ovoj vrsti korpusa. Svrha monitor-korpusa praćenje je promjena u jeziku, a njihov se sadržaj

redovito ažurira. Monitor-korpusi obavezno uključuju metapodatke o vremenu nastanka teksta. Primjer takvih korpusa prethodno je spomenuta skupina korpusa *Timestamped JSI Web Corpus*.

Posebna vrsta korpusa učenički su korpusi (eng. *learner corpus*), koji se sastoje od tekstova neizvornih govornika, a služe za proučavanje pogrešaka i problema na koje nailaze učenici stranih jezika. Primjeri učeničkih korpusa: za hrvatski jezik: CroLTeC (Mikelić Preradović 2020), za engleski jezik: *Cambridge Learner Corpus* (Nicholls 2003).

Korpusi se prema kriteriju označenosti dijele na sirove i označene korpuse. Sirovi korpusi (eng. *raw corpus*) sadržavaju samo tekst, tj. ne uključuju nikakve dodatne informacije o tekstu ili dodatne podatke umetnute u tekst. Većina je korpusa označena (eng. *annotated corpus*), što znači da sadržavaju podatke umetnute u tekst u formi oznaka. Na primjer, korpusi mogu sadržavati podatke o vrsti riječi svake pojavnice ili o pogreškama u tekstu. Bitno je naglasiti da se tekst i oznake moraju međusobno razlikovati, o čemu će biti više riječi u narednome poglavlju.

4.2.2.3. Označavanje korpusa

Budući da računala mogu obrađivati samo eksplicitne informacije, implicitne je značajke teksta potrebno učiniti eksplicitnima da bi ih računalo moglo obraditi. Proces pretvaranja implicitnih svojstava teksta u eksplicitna, pri čemu se tekst jasno i nedvojbeno razlikuje od oznaka, zove se kodiranje ili označavanje (eng. *annotation*, *encoding* ili neformalno *tagging*) (prema TEI Consortium 2021). Tekstovi sadržavaju mnogo implicitnih informacija na svim jezičnim razinama, ali i drugih vrsta informacija poput strukturnih i opisnih informacija. Na primjer, na morfološkoj razini svaka pojavnica teksta nosi implicitnu informaciju o vrsti riječi i drugim morfosintaktičkim značajkama. Rečenice, odlomci i naslovi nose strukturne informacije, dok se opisne informacije mogu odnositi na izvor i godinu izdavanja teksta.

Postoje razne sheme za označavanje korpusa. Neke od njih prihvaćene su u zajednici kao *de jure* ili *de facto* standardi. U nastavku slijedi prikaz dviju shema za označavanje kojima su označeni korpusi hrWaC i srWaC (korišteni u istraživanjima predstavljenima u narednim cjelinama knjige).

U sklopu projekta MULTEXT-East (*Multilingual Text Tools and Corpora for Central and Eastern European Languages*) (Erjavec 2012) razvijaju se korpusi i specifikacije za označavanje morfosintaktičkih značajki pojavnica za jezike središnje i istočne Europe te za engleski jezik kao *lingua franca*. Morfosintaktičke oznake (eng.

morphosyntactic description, MSD) organizirane su u kategorije, a svaka kategorija uključuje atribute koji pobliže opisuju tu kategoriju i njima odgovarajuće vrijednosti. Za hrvatski jezik postoji 13 kategorija, od kojih se prvih 10 odnosi na vrstu riječi, dok se preostale tri odnose na druge vrijednosti: imenica (N), glagol (V), pridjev (A), zamjenica (P), prilog (R), prijedlog (S), veznik (C), broj (M), čestica (Q), usklik (I); kratica (Y), ostalo (X) i interpunkcija (Z). Prvo slovo MSD-oznake odnosi se na kategoriju i može poprimiti jednu od navedenih 13 vrijednosti. Svaka od tih kategorija može sadržavati atribute s dodatnim MSD-informacijama. Na primjer, imenice sadržavaju atribute o vrsti imenica, rodu, broju, padežu i živosti (tim redoslijedom). Svaki atribut može poprimiti određenu vrijednost koja se bilježi malim slovom, dok se kategorije bilježe velikim slovom. Na primjer, prvi atribut o vrsti imenice može poprimiti vrijednost *c* za opću imenicu ili *p* za vlastitu imenicu, dok atribut o rodu može poprimiti vrijednosti *m* za muški, *f* za ženski ili *n* za srednji rod. Tako oznaka *Ncmsan* označava pojavnicu (npr. *naslon*) sa sljedećim MSD značajkama: *imenica, opća, muški rod, jednina, akuzativ, neživ*. Uz kategorije *usklici*, *kratice* i *interpunkcija* ne vezuju se dodatni atributi. Kategorijom *ostalo* označavaju se npr. strane riječi ili URL-ovi u tekstovima. O MULTEXT-East specifikacijama v. više na poveznici http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html.

Univerzalne ovisnosti (eng. *universal dependencies*, UD) međunarodni je projekt koji se bavi razvojem banaka stabala (eng. *treebank*), tj. sintaktički označenih korpusa za razne jezike, s ciljem provođenja dosljednoga prekojezičnoga označavanja, koje po potrebi dozvoljava jezično specifična proširenja skupa oznaka (Nivre et al. 2016). De Marneffe i Nivre (2019) ovisnosnu gramatiku (eng. *dependency grammar*) definiraju kao pristup sintaktičkoj analizi koji se zasniva na pretpostavci da se sintaktičke strukture sastoje prvenstveno od binarnih asimetričnih odnosa među riječima (*ovisnosni odnosi* ili *ovisnosti*). U ovisnosnoj se gramatici sintaktičke strukture opisuju pomoću ovisnosnoga stabla (eng. *dependency tree*) u kojemu čvorovi predstavljaju riječi, a strelice ili lukovi predstavljaju različite vrste ovisnosnih odnosa. Strukturni centar rečenice je predikat, dok su ostali čvorovi (tj. riječi) direktno ili indirektno povezani s njime ovisnosnim odnosima. O UD specifikacijama v. više na poveznici https://universaldependencies.org/. Ovom shemom označeni su noviji hrvatski korpusi ENGRI i ParlaMint-HR.

U sklopu projekta *Talkbank* (MacWhinney 2007), najvećega otvorenoga repozitorija prirodnojezičnih podataka govorenoga jezika, proširena je shema za anotiranje govorenoga jezika pod nazivom *Codes for Human Analysis of Transcripts* (*CHAT*) (MacWhinney 2000), koja je prvotno razvijena za označavanje dječjega govora. Ovaj je format kompatibilan s alatom *Computerized Language ANalysis* (*CLAN*), koji je razvijen za analizu prirodnojezičnih podataka transkribiranih u *CHAT* formatu. U sklopu programa za neke jezike moguće je izvršiti morfosintaktičko označavanje i

banku stabala, ali ta opcija ne postoji za hrvatski jezik. Shema pruža upute za transkribiranje diskursa za detaljnu fonološku i morfološku analizu. *CHAT* se ne temelji na XML-u (eng. *eXtensible Markup Language*), međutim postoji mogućnost prijevoda na taj format kako bi transkripti bili kompatibilni s drugim sustavima. Shema omogućuje segmentiranje transkripta na iskaze ili komunikacijske jedinice (eng. *communication unit*, *C-unit*), pohranjivanje raznih metapodataka u zaglavlju (npr. podatke o jezicima, sudionicima, lokaciji i sl.), označavanje pogrešaka i njihovu klasifikaciju itd. Anotacija transkripata i dalje se vrši ručno jer sustavi za prepoznavanje spontanoga govora nisu još dovoljno kvalitetni. Više o najnovijoj inačici uputa za upotrebu *CHAT* formata v. na poveznici: https://www.talkbank.org/manuals/CHAT.pdf.

*

U sklopu računalne pragmatike razvijaju se sheme za označavanje korpusa s pragmatičkom anotacijom. U nastavku slijedi prikaz dvaju primjera takvih shema prema Buntu (2009, 2017).

Shema *DIT++* temelji se na teoriji dinamičke interpretacije (eng. *dynamic interpretation theory*, *DIT*) – računalnome pristupu analizi značenja iskaza u dijalogu između čovjeka i čovjeka ili čovjeka i računala, koji je usmjeren na funkcionalne aspekte značenja iskaza. Prema teoriji dinamičke interpretacije govornik i recipijent iskazima ažuriraju svoja informacijska stanja. Pritom iskazi uz verbalna sredstva mogu uključivati i neverbalna sredstva (multimodalni iskazi). Shema *DIT++* sadržava 10 međusobno neovisnih dimenzija: (1) izvršavanje zadatka ili aktivnosti koja motivira dijalog; (2) povratna informacija o razumijevanju ili drugim aspektima obrade prethodnoga iskaza; (3) dijaloški činovi koje govornik koristi kako bi izrazio mišljenje o recipijentovoj obradi prethodnoga iskaza; (4) upravljanje preuzimanjem riječi; (5) upravljanje vremenom; (6) upravljanje kontaktom; (7) upravljanje vlastitom komunikacijom; (8) upravljanje komunikacijom sugovornika; (9) struktura diskursa; (10) upravljanje društvenim obavezama. Ova je shema temelj ISO-standarda pod oznakom HRN ISO 24617-2:2021 *Upravljanje jezičnim resursima – Okvir za semantičko označivanje (SemAF) – 2. dio: Dijaloški čin (ISO 24617-2:2020)* (v. više na poveznici https://repozitorij.hzn.hr/norm/HRN+ISO+24617-2%3A2021).

Prema Buntu (2017) dijaloški su činovi najproučavaniji pragmatički fenomen u računalnoj pragmatici, zbog kojega su u najvećoj mjeri razvijani korpusi s pragmatičkom anotacijom. Prvi pokušaj standardizacije sheme za označavanje višedimenzionalnih dijaloških činova shema je nazvana *Dialogue Act Markup Using Several Layers* ili *DAMSL* (Core i Allen 1997), koja nikada nije dovršena. Shema sadržava tri sloja: *komunikacijske funkcije usmjerene prema naprijed* (eng. *forward*

communicative function ili *forward-looking function*), *komunikacijske funkcije usmjerene prema natrag* (eng. *backward communicative function* ili *backward-looking function*) i *značajke iskaza* (eng. *utterance feature*). Svaki sloj ima skup dodatnih informacija koje mogu biti zabilježene za svaki iskaz. Funkcije usmjerene prema naprijed bilježe informacije koje se odnose na nastavak komunikacije (npr. postavljen zahtjev za informacijom), dok se funkcije usmjerene prema natrag odnose na skup informacija vezanih uz prethodni iskaz (npr. odgovor na postavljeno pitanje). Značajke iskaza opisuju sadržaj i strukturu iskaza. Prema Buntu (2017) nekoliko je primjera inačica *DAMSL* sheme koje su istraživači prilagodili potrebama svojih istraživanja, no shema je imala nedostatke poput nepreciznih definicija, loše raspodjele dimenzija te nepotpunoga inventara komunikacijskih funkcija.

4.2.2.4. Alati za razvoj i analizu korpusa

U ovome poglavlju slijedi prikaz alata za razvoj i analizu korpusa. U prvome dijelu poglavlja predstavljeni su skupovi alata za obradu hrvatskoga jezika *ReLDIanno* i *CLASSLA* te alati *Sketch Engine* i *NoSketch Engine*. U drugome dijelu slijedi prikaz izabranih alata za razvoj i anotaciju pragmatičkih korpusa i njihovih značajki.

Alat *ReLDIanno* (prema Ljubešić et al. 2013; Agić i Ljubešić 2015; Ljubešić i Erjavec 2016; Ljubešić et al. 2016; Ljubešić, Erjavec i Fišer 2016; Fišer, Ljubešić i Erjavec 2020) nudi mogućnost obrade hrvatskoga, slovenskoga i srpskoga jezika. Alat je moguće koristiti na dva načina: (1) preko *web*-aplikacije dostupne na poveznici http://clarin.si/services/web/, koja je primjerena za istraživače koji nisu tehnološki potkovani; (2) preko Python-biblioteke[67], koja je primjerena za tehnološki potkovane istraživače. Ovaj alat vrši segmentaciju teksta na rečenice. Sastavni je dio alata i tokenizator – poseban alat koji svodi tekst na pojavnice i vertikalizira ih. Vertikalizacija korpusa omogućuje njegovo označavanje dodatnim slojevima informacija za svaku pojavnicu. *ReLDIanno* nudi mogućnost vraćanja dijakritičkih znakova, a točnost te obrade za hrvatski jezik iznosi 99 %. Alat također nudi MSD označavanje oznakama MULTEXT-East, a točnost te obrade za hrvatski jezik iznosi 92,53 %. Osim toga, alat nudi mogućnost lematizacije (eng. *lemmatization*), odnosno dodjeljivanje leme svakoj pojavnici, pri čemu za hrvatski jezik postiže točnost od 99,5 %. Alatom je moguće izvršiti i označavanje imenovanih entiteta (eng. *named-entity recognition*, *NER*), odnosno automatsko označavanje pet kategorija: imena osoba i njihove izvedenice, lokacije, organizacije i 'razno'. Točnost alata pri označavanju vlastitih imena nije izračunata za hrvatski jezik. Posljednji sloj označavanja koje alat nudi sintaktička je raščlamba (eng. *parsing*)[68] primjenom

[67] Dostupna na poveznici https://github.com/clarinsi/reldi-lib.

[68] U hrvatskoj se literaturi još koriste sinonimi *parsanje* ili *parsiranje*.

univerzalne ovisnosti, a uključuje morfološku i sintaktičku anotaciju. Za evaluaciju sintaktičke raščlambe koriste se dvije mjere: mjera za neoznačene priključene vrijednosti (eng. *unlabeled attachment score*, *UAS*) i mjera za označene priključene vrijednosti (eng. *labeled attachment score*, *LAS*). Dok UAS-mjera uzima u obzir samo ovisnosni odnos, ali ne i semantičko obilježje priključeno toj ovisnosti, LAS-mjera zahtijeva da je ovisnosti ispravno priključeno njezino semantičko obilježje (Nivre i Fang 2017). Sintaktička raščlamba za hrvatski jezik postiže točnost oko 90 % za UAS-mjeru i oko 86 % za LAS-mjeru. Na *web*-aplikaciji moguće je izravno upisati tekst u predviđeno polje ili postaviti datoteke u sljedećim formatima: čiste tekstualne datoteke (*.txt*), MS Word-datoteke (*.doc*, *.docx*), PDF-datoteke (*.pdf*) i ZIP-format (*.zip*). Datoteke u ZIP-formatu mogu sadržavati kombinacije posljednjih triju dopuštenih tipova datoteka. Potrebno je odlučiti koje se obrade žele izvršiti na tekstu, a rezultati se vraćaju u trima različitim inačicama: tablica, JSON-format ili preuzimanje u datoteci TSV (eng. *tab-separated value*), u kojima su različite razine anotacija odvojene tabulatorom.

Tablica 5 prikaz je rezultata *web*-aplikacije *ReLDIanno* koja je obradila rečenicu *Hrvatska je članica Europske unije.* s opcijom MSD-oznaka, lematizacijom i NER-oznakama. Tablica 6 prikaz je rezultata obrade rečenice *Točnost alata nije izračunata za hrvatski jezik.* s opcijom MSD-oznaka, lematizacijom i UD-oznakama.

Alat *ReLDIanno* također nudi i *online*-inačicu flektivnoga leksikona hrvatskoga jezika, koji se može pretraživati prema površinskome obliku, lemi ili MSD-oznakama, a podržava i regularne izraze (eng. *regular expressions*, *regex*). O alatu v. više na poveznici https://www.clarin.si/info/k-centre/web-services-documentation/.

CLASSLA (*CLARIN Knowledge Centre for South Slavic Languages*) (prema Fišer, Ljubešić i Erjavec 2018; Ljubešić i Dobrovoljc 2019) su alati novije generacije koji nude mogućnost obrade hrvatskoga, srpskoga, slovenskoga, bugarskoga i makedonskoga jezika. Ova skupina alata nema *web*-aplikacije, već je dostupna isključivo preko Python-biblioteke[69]. Alat nudi mogućnost definiranja obrađuje li se standardni ili nestandardni hrvatski jezik. Alat vrši segmentaciju teksta na rečenice te uključuje i *ReLDI*-tokenizator. Nudi mogućnost MSD-označavanja primjenom oznaka MULTEXT-East i univerzalnih značajki UD-standarda, a točnost MSD-oznaka za standardni hrvatski jezik iznosi 94,18 % (što je više od alata *ReLDIanno*), a za nestandardni hrvatski jezik 95,11 %. Nadalje, alat nudi mogućnost lematizacije, a točnost za standardni hrvatski jezik iznosi 97,6 % (što je manje od alata *ReLDIanno*), dok za nestandardni hrvatski jezik iznosi 97,54 %. Alat nudi NER-oznake i sintaktičku raščlambu primjenom istih alata kao i alat *ReLDIanno*, a ove je opcije

[69] Dostupna na poveznici https://pypi.org/project/classla/.

moguće koristiti samo za obradu standardnoga jezika. O alatu v. više na poveznici https://www.clarin.si/info/k-centre/faq4croatian/.

	Surface	Tags	Lemma	Dep parse - gov / func
1.	Hrvatska	Npfsn	Hrvatska	B-loc
2.	je	Var3s	biti	O
3.	članica	Ncfsn	članica	O
4.	Europske	Agpfsgy	europski	B-org
5.	unije	Ncfsg	unija	B-org
6.	.	Z	.	O

Tablica 5. Primjer rezultata web-aplikacije *ReLDIanno* s MSD-oznakama, lematizacijom i NER-oznakama.

	Surface	Tags	Lemma	Dep parse - gov / func
1.	Točnost	Ncfsn	točnost	4 / nsubj
2.	alata	Ncmsg	alat	1 / nmod
3.	nije	Var3s	biti	4 / cop
4.	izračunata	Ncmpg	izračunati	0 / root
5.	za	Sa	za	7 / case
6.	hrvatski	Agpmsayn	hrvatski	7 / amod
7.	jezik	Ncmsan	jezik	4 / nmod
8.	.	Z	.	4 / punct

Tablica 6. Primjer rezultata web-aplikacije *ReLDIanno* s MSD-oznakama, lematizacijom i UD-oznakama.

*

Osim alata koji se primjenjuju pri izradi anotiranih korpusa, postoje alati koji mogu služiti i za njihovu analizu. Jedan od takvih alata komercijalni je alat *Sketch Engine*[70]. Za alat je inače potrebna pretplata, no u sklopu projekta ELEXIS (Krek et al. 2018) do ožujka 2022. godine besplatno je dostupan pristup pripadnicima hrvatske znanstvene zajednice prijavljivanjem preko AAI-identiteta.[71] Alatom *Sketch Engine* moguće je razviti i analizirati vlastite korpuse, ali i analizirati preko 500 korpusa na

[70] Dostupan na poveznici https://www.sketchengine.eu/.

[71] Sve informacije iz ovoga odlomka dostupne su na internetskim stranicama alata i u samome alatu.

više od 90 jezika koji se nalaze u sklopu alata. Alat je prvotno razvijen za leksikografsku korpusnu obradu prirodnojezičnih podataka, no u međuvremenu se razvio u alat koji osim leksikografa koriste i lingvisti, prevoditelji, učitelji i učenici stranih jezika, terminolozi, analitičari teksta i sl. U nastavku slijedi kratak prikaz značajki alata dostupnih za hrvatski jezik.

U sklopu alata *Sketch Engine* hrvatski su korpusi hrWaC, *Riznica*, CHILDES, OPUS2, DGT-MT i EUR-Lex, srpski korpus srWaC, korpusi OPUS2, *Timestamped JSI Web Corpus* i dr. Alat nudi mogućnost izrade skice riječi (eng. *word sketch*) na sažet način prikazujući kolokacije i ostale riječi u okolini riječi koja se istražuje, a služi za kratak prikaz njezinih gramatičkih i kolokacijskih značajki. Kolokacije su niz ili kombinacija riječi čije je supojavljivanje češće od očekivanoga slučajnoga supojavljivanja. Kolokati su grupirani, a pravila koja to definiraju navedena su u jezično ovisnim gramatikama skica. Osim grupiranja, u svakoj su grupi kolokati poredani od najveće prema najmanjoj vrijednosti mjere za izračun ovoga odnosa.[72] Alat također nudi mogućnost izrade usporedne skice riječi kada se žele usporediti dvije riječi. I ovdje su kolokati grupirani, i to na tri razine. Prva razina su kolokati koji se pojavljuju više uz prvu riječ nego uz drugu. Druga razina su kolokati koji se pojavljuju uz obje riječi. Treća razina su kolokati koji se pojavljuju više uz drugu nego uz prvu riječ. Dodana opcija alata tzv. je tezaurus koji nudi automatski generiran popis sinonima, antonima i sličnih riječi, a rezultat je također popis riječi poredan od najveće prema najmanjoj mjeri za izračun ovoga odnosa. Pomoću ovoga alata također je moguće pretraživati konkordancije, odnosno popise primjera riječi ili izraza u KWIC-formatu.

Key Word in Context ili KWIC-format prikazuje primjere upita koji su obojani crvenom bojom i nanizani jedan ispod drugoga, tako da se lijevi i desni kontekst jasno razlikuju. Kao upit mogu se postaviti riječi, izrazi, oznake (npr. MSD-oznake), dokumenti, vrste tekstova ili strukturni element korpusa – ovisno o razini anotacije pojedinoga korpusa. Napredne pretrage moguće je vršiti pomoću CQL-a (eng. *Corpus Query Language*), upitnoga jezika razvijenoga upravo za pretraživanje kompleksnih gramatičkih struktura ili leksičkih uzoraka u korpusima. U paralelnim je korpusima moguće vršiti pretraživanje paralelnih konkordancija. Ovaj alat nudi mogućnost generiranja popisa riječi, lema, MSD-oznaka i ostalih atributa s podatkom o frekvenciji pojavljivanja. Također je moguće generirati popis najfrekventnijih n-grama. N-gram je niz određenoga broja jedinica. Jedinice se mogu odnositi na slova, znamenke, pojavnice i sl., pa tako bigram sadržava dvije jedinice, trigram sadržava tri jedinice itd. U slučaju alata *Sketch Engine* n-grami se odnose na riječi, MSD-oznake ili leme. U konačnici, alat nudi mogućnost automatske ekstrakcije

[72] U sklopu knjige nećemo ulaziti u razne mjere koje se koriste u sklopu alata za izračune. V. više na https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf.

ključnih riječi i termina suprotstavljanjem promatranoga i referentnoga korpusa. Budući da se za ključne riječi i termine može reći da su tipični za promatrani korpus, jer se u tome korpusu pojavljuju češće nego u referentnome korpusu, alat kombinira statistiku i lingvističke kriterije za njihovu automatsku identifikaciju. Osim navedenih značajki alat nudi mogućnost izrade korpusa na temelju vlastitih dokumenata ili *web*-stranica. U sklopu alata nudi se automatska tokenizacija, lematizacija i MSD-označavanje za hrvatski jezik, a tehnologije u pozadini temelje se na starijim inačicama *ReLDIanno* alata. Moguće je izraditi vlastiti korpus spomenutim alatima *ReLDIanno* ili *CLASSLA*, a za analizu ga je moguće postaviti na alat *Sketch Engine*.

NoSketch Engine[73] inačica je otvorenoga koda alata *Sketch Engine* s ograničenim značajkama. Alat ne sadržava korpuse, već korisnici s odgovarajućim tehničkim znanjima mogu postaviti korpuse na svoje servere. Alat ne omogućava izradu skica riječi, tezaurus, izradu n-grama ni automatsku ekstrakciju terminologije. Ne nudi ni mogućnost izrade korpusa, no nudi napredno pretraživanje konkordancija i izradu popisa riječi. U sklopu ovoga alata nalaze se hrvatski korpusi hrWaC, *Riznica*, ParlaMint, ENGRI, DGT-MT i Tweet-hr te srpski korpus srWaC. Prednost ovoga alata u tome je što je besplatan i nema ograničenja na pretraživanje konkordancija i izradu popisa riječi koje ima *Sketch Engine*.

KonText[74] (prema Machálek 2014; Machálek 2020) još je jedan alat s vrlo sličnim specifikacijama kao *NoSketch Engine*, a na njemu se nalaze hrvatski korpusi hrWaC, *Riznica*, ParlaMint i ENGRI te srpski korpusi srWaC, Tweet-sr i *Torlak* (korpus torlačkoga narječja).

*

Za sada ne postoje razvijeni alati koji bi automatski mogli transkribirati i anotirati konverzacijske činove za potrebe računalnopragmatičkih istraživanja. Postojeća rješenja iz područja OPJ-a poput MSD-označavanja razvijena su na pisanim prirodnojezičnim podacima te nisu namijenjena (transkribiranome) govoru. Držimo da bi bilo zanimljivo provjeriti točnost alata *CLASSLA* za MSD-označavanje nestandardnih tekstova na HrAL-u. Zbog nekompatibilnosti formata i različitih tradicija bilježenja podataka takva provjera prvotno zahtijeva predobradu ulaznih podataka da bi ih alat uopće mogao obraditi. Nakon toga je potrebno izlazne podatke konvertirati za alate pomoću kojih će se provesti analiza. Potrebno je također izvršiti evaluaciju primjene alata na ovoj vrsti teksta te odvagati isplativost svih aktivnosti za dobiveni rezultat.

[73] https://nlp.fi.muni.cz/trac/noske

[74] https://www.clarin.si/kontext/corpora/corplist

Bunt (2017) napominje da se najviše računalnih programa u području računalne pragmatike razvija s ciljem podrške ručnoj anotaciji konverzacijskih prirodnojezičnih podataka. Za takve je korpuse karakteristična multimodalnost i višedimenzionalnost[75], što podrazumijeva označavanje različitih fenomena na različitim razinama koji se mogu međusobno preklapati (npr. istovremeno odvijanje kimanja glavom i iskaza „mmm" kao potvrde). Alati uglavnom nude pomoć pri povezivanju transkripta s audiosnimkom ili videosnimkom, segmentiranju transkripta na govorne činove, dodjeljivanju jedinstvenoga identifikacijskoga koda govornika govornim činovima, dodjeljivanju vremenskih oznaka za početak i kraj govornoga čina, dodjeljivanju različitih razina informacija, uvozu i izvozu podataka u različitim formatima i sl. (v. više u Müller i Strube 2006; Wittenburg et al. 2006; Carletta et al. 2009; Bunt, Kipp i Petukhova 2012).

Zanimljivo su istraživanje proveli Petukhova i Bunt (2011) primijenivši metode strojnoga učenja na klasifikaciju deset dimenzija sheme *DIT++*. Za dimenziju upravljanja komunikacije među sugovornicima postigli su točnost od gotovo 72 %, dok su za dimenziju govornikove procjene sugovornikove obrade prethodnoga iskaza postigli točnost od čak 96 %. Valja imati na umu da glavni izazov pri izradi korpusa govorenoga jezika predstavlja prikupljanje podataka, tj. snimanje govornika. Osim toga, korpusi korišteni u ovome istraživanju već su segmentirani na dijaloške činove, što je samo po sebi velik zadatak. Međutim, ovi rezultati daju nadu da se jedan takav alat može koristiti za poluautomatsko označavanje govornih činova prema shemi *DIT++* te time ubrzati barem ovaj korak obrade.

4.2.2.5. Korpusi hrvatskoga i srpskoga jezika

Koliko nam je poznato, *Korpus direktivnih govornih činova hrvatskoga jezika* (DirKorp) – koji je razvijen za potrebe istraživanja predstavljenoga u Cjelini 8 ove knjige – jedini je javno dostupan korpus hrvatskoga jezika s pragmatičkom anotacijom. Opis izrade DirKorp-a i njegovih obilježja dostupan je u Cjelini 9, a u ovome poglavlju slijedi prikaz hrvatskoga korpusa hrWaC te srpskoga korpusa srWaC, koji su korišteni za potrebe korpusnopragmatičkih istraživanja predstavljenih u Cjelinama 6 i 7. Osim toga, u poglavlju su prikazani i drugi odabrani korpusi hrvatskoga i srpskoga jezika.

Korpus hrWaC (prema Ljubešić i Klubička 2014: Ljubešić i Klubička 2016a) trenutačno je najveći korpus hrvatskoga jezika, veličine 1,4 milijarde pojavnica (v2.2). Razvijen je na Odsjeku za informacijske i komunikacijske znanosti Filozofskoga fakulteta Sveučilišta u Zagrebu. Sastoji se od *web*-dokumenata vršne

[75] Prikaz *DIT++* sheme v. u Poglavlju 4.2.2.3.

domene *.hr* prikupljenih 2011. i 2014. godine. Korpus je automatski označen na morfosintaktičkoj razini oznakama MULTEXT-East i lematiziran, a odlomci su izmiješani. Svaki odlomak sadržava metapodatke o URL-u, *web*-domeni i jeziku[76]. Korpus je besplatan i javno dostupan na repozitoriju CLARIN.SI[77] te u alatu otvorenoga koda *NoSketch Engine*[78]. U sklopu projekta ELEXIS (Krek et al. 2018) hrWaC je do ožujka 2022. godine pripadnicima hrvatske znanstvene zajednice besplatno dostupan i putem alata *Sketch Engine* prijavljivanjem preko AAI-identiteta.

Hrvatski nacionalni korpus ili HNK (prema Tadić 2002, Tadić 2007, Tadić 2009) reprezentativan je korpus veličine 234 milijuna pojavnica (v3.0). Razvijen je na Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. Sastoji se od tekstova napisanih na suvremenome hrvatskome standardnome jeziku objavljenih u razdoblju od 1990. godine. Riječ je o tekstovima različitih vrsta i žanrova o različitim temama. Korpus je automatski označen na morfosintaktičkoj razini oznakama MULTEXT-East i lematiziran. Zbog autorskih prava korpus nije javno dostupan, ali moguće je vršiti napredna pretraživanja preko alata *NoSketch Engine*.[79]

Hrvatski jezični korpus Riznica (prema Ćavar i Brozović Rončević 2012; Brozović Rončević et al. 2018) reprezentativan je korpus veličine 101 milijuna pojavnica (v0.1). Razvijen je na Institutu za hrvatski jezik i jezikoslovlje. Sastoji se od tekstova svih struka i funkcionalnih usmjerenja (npr. književnih, publicističkih i znanstvenih djela, osnovnoškolskih i srednjoškolskih udžbenika i sl.) koji uglavnom potječu iz razdoblja od druge polovice 19. stoljeća do danas. Korpus je automatski označen na morfosintaktičkoj razini oznakama MULTEXT-East i lematiziran. Sadržava metapodatke o autoru teksta, godini objavljivanja i sl.

Hrvatski korpus govornog jezika odraslih ili HrAL (Kuvač Kraljević i Hržica 2016) jedan je od rijetkih korpusa govornoga hrvatskoga jezika, a predstavlja reprezentativan uzorak govora neprofesionalnih odraslih izvornih govornika hrvatskoga jezika.

> HrAL je oblikovan uzorkovanjem spontane konverzacije između 617 govornika iz svih hrvatskih županija i sadrži više od 250.000 pojavnica i više od 100.000 različnica. Podatci su prikupljani u tri vremenska razdoblja: od 2010. do 2011., od 2014. do 2015. te tijekom 2016. godine. [...] Podaci su transkribirani, kodirani i segmentirani [...] (ibid.)

[76] hrWaC sadržava podatak o jeziku jer su se istovremeno prikupljali podaci za razvoj slovenskoga, srpskoga i bosanskoga *web*-korpusa.

[77] https://clarin.si/repository/xmlui/handle/11356/1064

[78] https://www.clarin.si/noske/run.cgi/corp_info?corpname=hrwac&struct_attr_stats=1

[79] http://filip.ffzg.hr/cgi-bin/run.cgi/first_form

Transkripti (165) su ručno anotirani *CHAT* formatom te sadržavaju podatke o dobi, spolu i stupnju obrazovanja govornika te o lokaciji, datumu, situaciji i trajanju konverzacije. Podaci o imenu i prezimenu i/ili inicijalima sudionika pseudoanonimizirani su (ibid.). Korpus je javno dostupan preko *TalkBanka*[80].

*

Korpus srWaC (prema Ljubešić i Erjavec 2011; Ljubešić i Klubička 2014b; Ljubešić i Klubička 2016b) trenutačno je najveći korpus srpskoga jezika, veličine više od 475 milijuna pojavnica (v1.2). Razvijen je na Odsjeku za informacijske i komunikacijske znanosti Filozofskoga fakulteta Sveučilišta u Zagrebu. Sastoji se od *web*-dokumenata vršne domene *.rs* prikupljanih 2014. godine. Korpus je automatski označen na morfosintaktičkoj razini oznakama MULTEXT-East i lematiziran, a odlomci su izmiješani. Svaki odlomak sadržava metapodatke o URL-u, *web*-domeni i jeziku[81]. Korpus je besplatan i javno dostupan na repozitoriju CLARIN.SI[82] i u alatu otvorenoga koda *NoSketch Engine*[83]. U sklopu projekta ELEXIS (Krek et al. 2018) srWaC je do ožujka 2022. godine pripadnicima hrvatske znanstvene zajednice besplatno dostupan i putem alata *Sketch Engine* prijavljivanjem preko AAI-identiteta.

Korpus suvremenog srpskog jezika ili SrpKor2013 (prema Krstev i Vitas 2005, Vitas i Krstev 2012) veličine je oko 122 milijuna pojavnica. Razvijen je na Matematičkome fakultetu Univerziteta u Beogradu. Budući da su tekstovi korpusa zakodirani kodnom stranicom ASCII, korpus je zakodiran kodnom shemom *aurora*[84] (npr. umjesto znaka *Č* koristi se kombinacija znakova *CY* ili *Cy*, a umjesto znaka *č* koristi se kombinacija znakova *cy*). Korpus se sastoji od tekstova različitih vrsta i žanrova o različitim temama iz 20. i 21. stoljeća. Na razini dokumenta postoji informacija o funkcionalnome stilu na kojemu je napisan tekst. Korpus je automatski označen na razini vrsta riječi primjenom vlastite sheme i lematiziran. Pristup korpusu omogućen je slanjem maila održavateljima korpusa.

[80] https://sla.talkbank.org/TBB/ca/Croatian

[81] srWaC sadržava podatak o jeziku jer su se istovremeno prikupljali podaci za razvoj slovenskoga, srpskoga i bosanskoga *web*-korpusa.

[82] https://www.clarin.si/repository/xmlui/handle/11356/1063

[83] https://www.clarin.si/noske/run.cgi/corp_info?corpname=srwac&struct_attr_stats=1

[84] V. više na poveznici http://www.korpus.matf.bg.ac.rs/prezentacija/uputstvo.html#aurora.

4.2.2.6. Primjeri pragmatičkih korpusa

Broj velikih korpusa sa sustavno provedenom pragmatičkom anotacijom za sada je malen. Zbog nerazmjernoga odnosa između pragmatičkih funkcija i sredstava (formi) njihova izražavanja automatska anotacija korpusa nije moguća. Iz toga se razloga tek malen broj istraživača upustio u izradu većih korpusa ove vrste. Uglavnom se za potrebe korpusnopragmatičkih istraživanja izrađuju specijalizirani korpusi manjega obima namijenjeni pojedinim istraživanjima (v. prikaz obilježja i izrade specijaliziranoga korpusa govornih činova DirKorp u Cjelini 9). Osim toga, pragmatička se istraživanja nekada provode i na korpusima bez pragmatičke anotacije.

Primjer korpusa koji ne sadržava pragmatičku anotaciju, ali na kojemu se provode pragmatička istraživanja jest *Birmingham Blog Corpus* (prema Kehoe i Gee 2007; Kehoe i Gee 2012). To je ustvari potkorpus većega skupa korpusa razvijenih na odjelu *Research and Development Unit for English Studies* na sveučilištu *Birmingham City University*. Sastoji se od objava na blogovima i čitateljskih komentara veličine 600 milijuna pojavnica engleskoga jezika koji su prikupljani u razdoblju od 2000. do 2010. godine. Korpus je automatski označen na razni vrste riječi primjenom vlastite sheme[85], a dokumenti sadržavaju metapodatak o datumu objave. Pristup korpusu[86] omogućen je slanjem maila održavateljima korpusa. Lutzky i Kehoe (2016) na ovome su korpusu proveli istraživanje o identifikaciji psovki uzimajući u obzir kontekst u slučaju višeznačnih leksema. Isti su istraživači proveli i istraživanja o isprikama (Lutzky i Kehoe 2017a; Lutzky i Kehoe 2017b) kao govornim činovima koji sadržavaju formulaične izraze, što omogućava njihovo lakše pretraživanje u korpusima primjenom dostupnih alata. Primjenom popisa markera ilokucijske snage koji se upotrebljavaju u isprikama te pretraživanjem kolokacija prikazali su kako precizirati rezultate upita i smanjiti ručnu analizu koja je inače potrebna za analizu rezultata upita korpusa. Zanimljivo je i kontrastivno istraživanje koje je provela Popoola (2017) na temu utvrđivanja značenja naziva brendova za potrebe razrješavanja sporova oko zaštitnih znakova proizvođača. U studiji je predstavljena metoda sintaktičko-pragmatičke analize korpusa kao alternative rječnicima i anketama tržišta. U istraživanju su korišteni sljedeći korpusi: *Birmingham Blog Corpus*, jedan korpus novinskih članaka i jedan korpus sastavljen od objava na Twitteru (ibid.). Slična pragmatička istraživanja provode se i na drugim korpusima bez pragmatičke anotacije. U takvim slučajevima istraživači osmišljavaju metode izoliranja podataka potrebnih za proučavanje pragmatičkih fenomena. Takva istraživanja provedena su u sklopu ove knjige i predstavljena u Cjelinama 6 i 7.

[85] V. više o POS-oznakama v. na poveznici http://wse1.webcorp.org.uk/guide/tagsets.html.

[86] Dostupno na poveznici http://wse1.webcorp.org.uk/cgi-bin/BLOG/index.cgi.

Kada je riječ o javno dostupnim hrvatskim i srpskim korpusima, koliko nam je poznato, za sada ne postoje korpusi govorenoga jezika s gramatičkom i/ili pragmatičkom anotacijom, kao ni korpusi pisanoga jezika s pragmatičkom anotacijom. Zato u nastavku poglavlja slijedi prikaz izabranih pragmatičkih korpusa engleskoga i drugih stranih jezika.

Prema Buntu (2017) većina postojećih korpusa s pragmatičkom anotacijom sadržava oznake o diskursnim odnosima u pisanim tekstovima te o govorenim dijaloškim činovima. Primjer jednoga takvoga većega korpusa jest *Penn Discourse Treebank* ili PDTB (Prasad, Webber i Lee 2018), koji sadržava oznake o diskursnim odnosima, odnosno strukturi diskursa i njegovoj semantici. U sklopu većega korpusa *Penn Treebank* (PTB) dodane su diskursne anotacije na dio tekstova objavljenih u novinama *Wall Street Journal* veličine 1 milijun pojavnica. Anotacije diskursnih odnosa teorijski su neutralne jer nisu bilježene ovisnosti među odnosima.[87] Bunt (2017) navodi da postoje korpusi drugih jezika razvijeni za potrebe proučavanja supojavljivanja diskursnih oznaka, kao što su kineski, češki, nizozemski, njemački, hindi i turski – uz napomenu da su ti korpusi ručno označeni i skromnoga obujma. Osim toga, za svaki je korpus razvijena zasebna shema zasnovana na različitim teorijskim polazištima.

DialogBank (prema Bunt et al. 2019) jedan je od rijetkih dijaloških korpusa koji su označeni standardom ISO 24617-2 (v. Cjelinu 4.2.2). Razvijen je na sveučilištu *Tilburg University*, a sastoji se od postojećih dijaloških korpusa anotiranih drugim shemama. Od toga su četiri korpusa engleskoga jezika (*HCRC Map Task*[88], *Switchboard*[89], TRAINS[90] i DBOX[91]) i četiri korpusa nizozemskoga jezika (DIAMOND[92], OVIS[93], *Dutch Map Task*[94] i *Schiphol*[95]). U nekim slučajevima zadržane su originalne anotacije kako bi se mogle usporediti sheme za označavanje. O DialogBanku v. više na poveznici https://dialogbank.uvt.nl/.

Sljedeći primjer korpusa s pragmatičkom anotacijom korpus je *Engineering Lecture Corpus* (prema Alsop i Nesi 2013; Alsop i Nesi 2014), koji sadržava 76 transkripata. Transkripti se temelje na videosnimkama u trajanju od otprilike jedan sat. Riječ je o

[87] V. više na poveznici https://www.seas.upenn.edu/~pdtb/.

[88] V. više u Anderson et al. (1991).

[89] V. više u Godfrey, Holliman i McDaniel (1992).

[90] V. više u Allen et al. (1995).

[91] V. više u Petukhova et al. (2014).

[92] V. više u Geertzen et al. (2004).

[93] V. više na poveznici http://www.let.rug.nl/vannoord/Ovis/.

[94] V. više u Caspers et al. (2000).

[95] V. više u Prüst, Minnen i Beun (1984).

snimkama predavanja održanih na engleskome jeziku na trima sveučilištima: *Coventry University* u Velikoj Britaniji, *Universiti Teknologi* u Maleziji i *Auckland University of Technology* na Novome Zelandu. Korpus je razvijen na sveučilištu Coventry University s trima ciljevima: za potrebe utvrđivanja i opisivanja tipičnih značajki diskursa inženjerskih predavanja; za potrebe uspoređivanja stilova inženjerskih predavanja na engleskome jeziku u različitim dijelovima svijeta te općenito radi razvoja i testiranja sustava pragmatičke anotacije za korpuse. Korpus pokriva područja građevinarstva, strojarstva i elektrotehnike. Shema korpusa temelji se na XML-u i pridržava se standarda strukturnih oznaka definiranih u Smjernicama TEI[96]. Više riječi o Smjernicama TEI bit će u Cjelini 9. Za svakoga je govornika zabilježena informacija o spolu i akademskome statusu. Od metapodataka svaki transkript sadržava opis datoteke (uključujući naslov, opis izvora snimke i informacije o transkripciji), opis kodiranja te ostale informacije (npr. o govornicima, značenju identifikatorskih oznaka i sl.). Oznake su unesene ručno na temelju videosnimki, a za sve su oznake provjereni koeficijenti pouzdanosti anotacija među anotatorima (utvrđeno je koliko su često anotatori donijeli istu odluku). Korpus sadržava oznake triju pragmatičkih značajki: humor, pripovijedanje i sažimanje.[97] Svaka značajka definirana je jednim XML-elementom i može sadržavati jedan od atributa. Svaki atribut uključuje dodatne informacije s pobližim opisom značajki: (1) atributi za značajku 'humor': *vulgaran*, *crni*, *uvredljiv*, *razigran*, *samoomalovažavajući*, *ironija/sarkazam*, *vic*, *zadirkivanje* te *igra riječima*; (2) atributi za značajku 'pripovijedanje': *priča*, *prepričavanje*, *anegdota* i *oprimjerivanje*; (3) atributi za značajku 'sažimanje': *pregled sadržaja prethodnoga predavanja*, *pregled sadržaja trenutnoga predavanja*, *najava sadržaja trenutnoga predavanja* te *najava sadržaja budućega predavanja*. U korpusu su također označene pauze, smijeh, pisanje ili crtanje na ploči i sl. Korpus nije javno dostupan, kao ni upute za dobivanje pristupa korpusu. O korpusu v. više na poveznici www.coventry.ac.uk/elc.

Korpus SPICE-Ireland[98] (skraćeno od *Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland*) sastavnica je korpusa *International Corpus of English: Ireland Component* (ICE-Ireland) s anotacijom pragmatičkih, diskursnih i prozodijskih obilježja. SPICE-Ireland uključuje uzorke različitih vrsta privatnih i javnih, formalnih i neformalnih dijaloga i monologa opsega od oko 2000 riječi. Korpus sadržava ukupno 626 597 riječi, a članovima akademske zajednice dostupan je za besplatnu upotrebu na zahtjev autorima. ICE-Ireland korpus je govorenoga i pisanoga standardnoga engleskoga jezika, a nastao je u okviru projekta izrade

[96] Smjernice TEI razvija i održava konzorcij TEI (*Text Encoding Initiative*), a namijenjene su svima koji se bave pripremom i/ili obradom tekstualnih resursa u digitalnome obliku.

[97] Preuzeto s poveznice https://www.coventry.ac.uk/research/research-directories/current-projects/2015/engineering-lecture-corpus-elc/annotations-and-mark-ups/.

[98] Podaci o korpusu SPICE-Ireland koji se navode u ovome poglavlju preuzeti su iz priručnika Kallena i Kirka *SPICE-Ireland: A User's Guide* (2012).

jezičnih korpusa u zemljama u kojima engleski ima status prvoga ili službenoga jezika. Jedan od glavnih ciljeva izrade ovih korpusa proučavanje je prekograničnih jezičnih varijacija između Irske i Sjeverne Irske – kako na gramatičkoj i leksičkoj razini (ICE-Ireland), tako i na razini jezičnoga ponašanja (SPICE-Ireland). Pod jezičnim se ponašanjem pritom podrazumijevaju prozodijska obilježja iskaza te upotreba raznih interaktivnih elemenata govorenoga diskursa.

Transkripti usmenih komunikacijskih činova koje SPICE-Ireland obuhvaća izvorna su sastavnica korpusa ICE-Ireland. Jezična građa korpusa SPICE-Ireland sastoji se od uzoraka 15 diskursnih tipova usmene komunikacije (konverzacija licem u lice, telefonski razgovori, debate u parlamentu, transkripti sudskih ispitivanja, izjave, prezentacije i izvještavanje u medijima itd.) prikupljene od 945 izvornih govornika. U korpusu su uz svaki tekst navedeni podaci o mjestu i vremenu nastanka transkripta, naslov/tema teksta te demografski i psihografski podaci o govornicima. U korpusu je zastupljena podjednaka količina jezičnoga materijala s područja Republike Irske i Sjeverne Irske. Korpus je popraćen priručnikom za korisnike u kojemu su detaljno prikazana sva njegova obilježja, s posebnim naglaskom na pravilima transkripcije i anotacije jezične građe.

Korpus SPICE-Ireland uključuje anotaciju prozodije, vrsta govornih činova, diskursnih markera, citata i drugih pragmatičkih pojavnosti. Pragmatička anotacija u ovome korpusu uključuje markiranje govornih činova u skladu sa Searleovom klasifikacijom na reprezentative, direktive, komisive, ekspresive i deklarative (usp. Poglavlje 1.3.3). Govorni činovi koji nisu upotrijebljeni u doslovnome značenju (kao modulacije primarnih govornih činova) označeni su posebnom oznakom. U skladu s dogovorom da bi svaki iskaz trebao biti klasificiran prema vrsti govornoga čina, autori korpusa uveli su posebnu kategoriju i oznaku za one elemente govora koji su važni za održavanje (kontinuiteta) diskursa, ali nemaju jasno definiranu funkciju kao govorni činovi. Nadalje, u korpusu su posebno markirani izrazi s interakcijskom funkcijom (eng. *social expressions*) koji nemaju puno propozicijsko značenje te nisu povezani s održavanjem (kontinuiteta) diskursa (npr. pozdravi i čestitke). Uvedena je oznaka za one elemente koji se ne mogu analizirati na pragmatičkoj razini, odnosno koji se ne mogu klasificirati kao govorni činovi ili sastavnice konverzacijskoga čina (najčešće zbog svoje nepotpunosti ili dvosmislenosti). Uz pragmatičku anotaciju SPICE-Ireland ima i prozodijsku anotaciju koja se odnosi na snagu, glasnoću, kvalitetu i trajanje glasa, a predviđena je za potrebe proučavanja odnosa između prozodije, sintakse i pragmatike. Nadalje, u korpusu su zasebno anotirane diskursne oznake kao elementi diskursa koji označavaju govornikov odnos prema ilokucijskoj jezgri iskaza, odnos između govornika i sugovornika, promjenu teme, pojašnjavanje iskaza itd. Pritom vokalni elementi u korpusu čija upotreba nije povezana s ilokucijskom snagom iskaza (npr. poštapalice) nisu klasificirani kao diskursne

oznake. Diskursne oznake u korpusu podijeljene su u tri skupine: sintaktičke, leksičke i fonološke. Kategorija sintaktičkih diskursnih markera u korpusu uključuje subjekte *I* i *you* u kombinaciji s glagolima percepcije, govora ili znanja. Leksički diskursni markeri najbrojnija su vrsta diskursnih markera u korpusu, a sastoje se od jedne ili više riječi. Fonološki diskursni markeri skupine su glasova u funkciji diskursnoga markera koji nemaju status leksema u leksikonu engleskoga jezika. U korpusu su anotirani i citatni markeri (glagoli, čestice i njihove kombinacije kojima govornik najavljuje citat) te rečenične oznake (skupina izraza koji se javljaju na kraju rečenice, a koji su u pravilu deiktični te upućuju na neki element u diskursu (prethodno iznesene informacije, govornika, sugovornika i sl.). Kao zasebna podskupina rečeničnih oznaka izdvajaju se vokativne oznake na finalnoj poziciji iskaza.

Iz ovoga opisa korpusa SPICE-Ireland vidljivo je da se provedena pragmatička anotacija odnosi na više vrsta pragmatičkih informacija. Međutim, njome nisu obuhvaćene sve pragmatičke kategorije, stoga valja imati na umu da nije svaki korpus s pragmatičkom anotacijom pogodan za svako pragmatičko istraživanje. Primjerice, istraživač zainteresiran za proučavanje poštapalica ne bi bio u mogućnosti provesti vertikalnu analizu njihove upotrebe (*funkcija* → *forma*) jer one nisu označene u ovome korpusu. Međutim, u slučaju proučavanja vrsta govornih činova, izraza s interakcijskom funkcijom, deiktika ili diskursnih markera, korpus SPICE-Ireland istraživaču bi omogućio njihovo pretraživanje te stjecanje uvida koje nijedna druga istraživačka metoda ne omogućava.

Primjer pragmatičke anotacije u korpusu SPICE-Ireland (zajedno s drugim korpusima ove vrste) može poslužiti kao vrijedan uzor za izradu korpusa hrvatskoga jezika s pragmatičkom anotacijom, ali i kao uzor za izradu manjih specijaliziranih korpusa za potrebe pojedinih pragmatičkih istraživanja.

4.2.3. Korpusnopragmatički pristup jeziku

Korpusna pragmatika počinje se razvijati početkom 21. stoljeća kao interdisciplinarno područje koje povezuje lingvističku pragmatiku i računarstvo, a usmjereno je na izradu računalnih prirodnojezičnih korpusa te na njihovu primjenu za potrebe proučavanja pragmatičkih fenomena u pisanome i govorenome jeziku. Korpusni pristup jeziku lingvisti su dugo vremena smatrali nespojivim s pragmatikom (Romero-Trillo 2008: 2). Naime, dok korpusni pristup jeziku podrazumijeva obradu autentične jezične građe primjenom pedantno razrađenih kvantitativnih istraživačkih metoda, pragmatička su istraživanja i danas pretežno kvalitativnoga tipa – zasnovana na istraživačevoj introspekciji, podacima dobivenim metodom elicitacije ili analizom

autentične jezične građe maloga obima. Primjena korpusne analize u istraživanjima pragmatičkih fenomena predstavlja velik obrat u razvoju pragmatike, prije svega zbog toga što omogućava sustavnu analizu autentične jezične građe velikoga obima, a time i otkrivanje obrazaca jezične upotrebe koji kvalitativnim analizama „prolaze ispod radara" (ibid.). Osim toga, valja naglasiti da primjena novih tehnologija u lingvistici, pa tako i pragmatici, nije samo omogućila ili olakšala/ubrzala brojne istraživačke procese, nego je otvorila vrata novom, drugačijem načinu razmišljanja o jeziku (Leech 1992a).

> Upotreba velikih korpusa, podržanih softverskim alatima za pretraživanje podataka, omogućava sustavno provođenje empirijski zasnovanih pragmatičkih istraživanja. To ide u prilog razvoju validnijih pragmatičkih teorija šireg dometa. Provođenje korpusnih istraživanja može rezultirati manjim izmjenama postojećih teorija, no može dovesti i do potpunoga preosmišljavanja pragmatičkih koncepata i teorijskih okvira. (Bunt 2017: 327)

Same začetke korpusne pragmatike Aijmer i Rühlemann (2015) smještaju u 2004. godinu, kada je objavljeno posebno izdanje časopisa *Journal of Pragmatics* posvećeno korpusnoj lingvistici (ur. Jacob Mey). Od tada do danas objavljena je nekolicina zbornika i udžbenika posvećenih korpusnoj pragmatici, kojima su postavljeni njezini temelji (v. npr. Adolphs 2008; Romero-Trillo 2008, *Yearbooks of Corpus Linguistics and Pragmatics* 2013–2016; Jucker, Schreier i Hundt 2009; Felder et al. 2011, Aijmer i Rühlemann 2015, Rühlermann 2018). Korpusnopragmatičke konferencije počele su se održavati krajem prvoga desetljeća 21. stoljeća (prva je održana 2007. godine u Švedskoj pod naslovom *Pragmatics, Corpora and Computational Linguistics*) te se i dalje održavaju u organizaciji raznih institucija diljem svijeta. Godine 2017. pokrenut je korpusnopragmatički časopis *Corpus Pragmatics: International Journal of Corpus Linguistics and Pragmatics.* Danas se korpusna pragmatika predaje i kao zaseban predmet u okviru studijskih programa lingvistike na svjetskim sveučilištima, dok se na hrvatskim sveučilištima za sada predaje u sklopu kolegija posvećenih pragmatici i korpusnoj lingvistici.

*

U pionirskome udžbeniku iz korpusne pragmatike *Corpus Pragmatics: A Handbook* (2015) autori Aijmer i Rühlerman ovu disciplinu definiraju kao spoj pragmatičkoga pristupa jeziku (koji podrazumijeva kvalitativnu, horizontalnu analizu) i korpusnoga pristupa jeziku (koji podrazumijeva kvantitativnu, vertikalnu analizu). Pragmatički pristup u pravilu je kvalitativan (horizontalan) jer je usmjeren na sintagmatske odnose među sastavnicama iskaza/diskursa te na njihove pragmatičke funkcije. S obzirom na kontekstualnu uvjetovanost jezika u upotrebi pragmatička se istraživanja

obično provode na manjemu broju tekstova nad kojima je moguće „ručno" provesti horizontalnu analizu jezične građe uzimajući u obzir kontekstualne faktore:

> Zbog usmjerenosti na pojedinačne tekstove pragmatička su istraživanja u suštini kvalitativna: fokus nije na broju pojavnica već na funkcijama jezičnih jedinica u tekstovima koji se analiziraju. S obzirom na zavisnost od konteksta, pragmatička se istraživanja metodološki zasnivaju na analizi manjega broja tekstova, koje je moguće podvrgnuti detaljnoj horizontalnoj analizi. Pritom se veliki i često cjeloviti tekstovi interpretiraju u istome vremenskom slijedu u kojemu su producirani i percipirani. Ova metodologija […] snažno odudara od 'vertikalne' metodologije koja je prevalentna u korpusnoj lingvistici (ibid. 3)

Korpusni pristup jeziku podrazumijeva vertikalnu analizu (prethodno obrađene) jezične građe, koja je usmjerena na utvrđivanje paradigmatskih jezičnih odnosa i frekventnosti pojedinih jezičnih pojava u jezičnoj upotrebi. Rezultat korpusnolingvističke vertikalne analize u pravilu je neka vrsta frekvencijskoga prikaza upotrebe i obilježja jezičnih jedinica koje su predmet istraživanja (ibid.).

Korpusna pragmatika obično kombinira vertikalnu i horizontalnu jezičnu analizu, i to primjenom dviju mogućih metoda (ibid. 9):

> (1) **Metoda 1** (*forma → funkcija*) podrazumijeva istraživanje koje se vrši provođenjem vertikalne, a potom horizontalne analize. Prvo se u korpusu (vertikalno) pretražuje određena jezična forma (riječ ili sintagma koja vrši neku pragmatičku funkciju), a zatim se analiziraju funkcije njezinih pojavnica u korpusu.

Primjeri korpusnopragmatičkih istraživanja u kojima je primijenjena Metoda 1 dostupni su u Cjelini 6 (korpusnopragmatička analiza pragmatičkih funkcija i obilježja određenih i neodređenih pridjeva) i Cjelini 7 ove knjige (korpusnopragmatička analiza pragmatičkih funkcija i obilježja glagola u imperativu). U prvoj fazi dvaju istraživanja izvršena je vertikalna analiza koja uključuje pretragu korpusa hrWaC i srWaC te izradu uzoraka KWIC-primjera određenih i neodređenih pridjeva (Cjelina 6) i glagola u imperativu (Cjelina 7) u odgovarajućoj formi. U drugoj fazi istraživanja provedena je horizontalna analiza pojedinih primjera na temelju koje je izvršena anotacija oznakama koje se odnose na njihove pragmatičke i funkcije i obilježja. Potom je uslijedila statistička obrada podataka te interpretacija dobivenih rezultata.

> (2) **Metoda 2** (*funkcija → forma*) podrazumijeva obratno postavljenu analizu u odnosu na Metodu 1: korpus se pretražuje prema pragmatičkoj funkciji koju vrše određene riječi ili veće jezične jedinice (npr. performativi).

Primjenu Metode 2 omogućava primjerice upotreba našega specijaliziranoga korpusa direktivnih govornih činova hrvatskoga jezika DirKorp. Ovaj korpus omogućava provođenje vertikalne analize jer nudi mogućnost pretraživanja govornih činova s obzirom na njihove pragmatičke funkcije i obilježja (usp. Cjeline 8 i 9).

Jedan od najvećih metodoloških problema s kojima se korpusni pragmatičari suočavaju, ističu Aijmer i Rühlerman (ibid. 10), nerazmjeran je odnos između pragmatičkih funkcija i jezičnih sredstava (formi) kojima se te funkcije izražavaju. Jedna forma može vršiti više pragmatičkih funkcija u diskursu, kao što se jedna funkcija može izražavati različitim formama, što znatno otežava proces pretraživanja korpusa prema kriteriju pragmatičke funkcije. Upravo iz toga razloga korpusni pragmatičari najčešće istražuju konvencionalizirane govorne činove, odnosno funkcije koje se izvršavaju ograničenim brojem jezičnih sredstava (Jucker, Scheier i Hundt 2009: 3). Primjerice, funkcija pozdravljanja sugovornika može se izvršiti ograničenim brojem jezičnih sredstava u nekome jeziku, stoga je ta tema pogodna za proučavanje primjenom korpusnopragmatičkoga pristupa. U slučaju nekonvencionaliziranih govornih činova i pragmatičkih funkcija koje se izražavaju višestrukim jezičnim sredstvima (npr. vrlo je široka paleta jezičnih jedinica koje mogu vršiti funkciju poštapalice) nužna je pragmatička anotacija korpusa kao preduvjet za istraživanje odnosa pragmatičke funkcije i sredstva njezina izražavanja. Prvo se redom označavaju pragmatičke funkcije jezičnih jedinica u diskursu, a potom se pragmatički anotiran korpus pretražuje prema oznakama za funkciju (Aijmer i Rühlerman 2015: 10). Ovakvi su korpusi u pravilu manjega obima jer zahtijevaju ručno anotiranje.

Zbog velikoga opsega i razgranatosti područja bavljenja pragmatike, metodologije korpusnopragmatičkih istraživanja ne mogu se svesti pod zajednički nazivnik,[99] već svako od jezgrenih područja pragmatike (deiksa, referencijalnost, govorni činovi, implikatura, presupozicija, konverzacijska analiza i dr.), pa tako i svaka pojedinačna tema zahtijeva razradu specifične metodologije u skladu s predmetom i ciljevima istraživanja.[100]

U Cjelinama 6, 7 i 8 ove knjige slijedi prikaz primjera korpusnopragmatičkih (kontrastivnih) analiza pojedinih gramatičko-pragmatičkih kategorija hrvatskoga i srpskoga jezika. Svaka analiza ima specifične metodološke postavke prilagođene predmetu istraživanja te vrsti korpusa koji se analizira. Opće postavke triju korpusnopragmatičkih analiza prikazane su u Cjelini 5, koja je posvećena korpusnopragmatičkome pristupu gramatičko-pragmatičkih pojavnosti u jeziku u

[99] Osim načelne podjele na Metodu 1, Metodu 2 ili njihovu kombinaciju.

[100] Razrade i primjeri primjene korpusnopragmatičke metode u okviru pojedinih područja pragmatike dostupne su u prethodno spomenutim korpusnopragmatičkim udžbenicima i zbornicima, kao i u nastavku ove knjige.

upotrebi. Cjelina 9 posvećena je prikazu izrade specijaliziranoga korpusa direktivnih govornih činova koji smo izradili za potrebe istraživanja prikazanoga u Cjelini 8.

4.2.4. Računalnopragmatički pristupi interpretaciji i generiranju govornih činova

Formalizacija inferencijskih procesa u međuljudskoj komunikaciji od samih začetaka OPJ-a predstavlja jedan od njezinih najvećih izazova (Bunt i Black 2000). U pragmatici pojam *inferencija* odnosi se na interpretativni proces zaključivanja o tome kako se (doslovno) značenje iskaza razlikuje od njegova smisla (Tomlinson i Bott 2013: 3569), tj. na proces prepoznavanja značenja koje govornik slušatelju prenosi putem iskaza na implicitnoj razini. Da bi slušatelj mogao uspješno donijeti zaključak o takvome implicitnome značenju, treba se osloniti na informacije koje nisu eksplicitno iskazane u iskazu, već su dio njegova i govornikova zajedničkoga znanja (usp. Poglavlje 1.3.2).

Jedno od područja bavljenja računalne pragmatike razvoj je modela i algoritama koji „popunjavaju" informacije koje nisu eksplicitno iskazane u iskazu. Četiri su temeljna računalnopragmatička inferencijska problema: (1) razrješavanje (ko)referencijalnosti; (2) interpretacija i generiranje govornih činova; (3) interpretacija i generiranje diskursne strukture i koherentnih sveza; (4) abdukcija (usp. Poglavlje 4.2.1). U sklopu ovoga poglavlja predstavit ćemo temeljne izazove interpretacije govornih činova kao primjer jednoga od inferencijskih računalnopragmatičkih problema.

Pregled temeljnih izazova interpretacije govornih činova koji slijedi u nastavku poglavlja temelji se na radu Danijela Jurafskyja *Pragmatics and Computational Linguistics* (2004), jednoga od autora temeljnoga udžbenika iz područja OPJ-a (Jurafsky i Martin 2009).

Problem interpretacije govornih činova odnosi se na utvrđivanje vrste govornoga čina koji se ostvaruje nekim iskazom. Neke govorne činove lako je identificirati jer imaju površinski oblik. Npr. govorni čin postavljanja pitanja, ako je eksplicitno iskazan, sastoji se upitnoga iskaza (rečenice) koji može uključivati upitnu zamjenicu ili česticu *li,* a u pisanim tekstovima u pravilu je označen upitnikom. Iz površinskih oblika moguće je razviti apstraktna strukturna obilježja iskaza. Međutim, mnoge vrste govornih činova nemaju jasno definirana strukturna obilježja. Primjerice, govorni čin molbe može se ostvariti iznošenjem imperativnoga iskaza (Primjer 37a), deklarativnoga iskaza (tvrdnje) (Primjer 37b) ili upitnoga iskaza (Primjer 37c).

Primjer 37

(a) Dodaj mi vode (molim te)!
(b) Molim te da mi dodaš vode.
(c) Hoćeš li mi (molim te) dodati vode?

Interpretacija govornih činova u računalnoj se pragmatici zasniva na dva pristupa: na logički utemeljenome pristupu te na pristupu utemeljenome na vjerojatnosti. Tvorci logički utemeljenoga pristupa ili inferencijskoga pristupa (eng. *inferential approach*) jesu Gordon i Lakoff (1971) te Searle (1975) (prema Jurafsky 2004). Shodno autorima prilikom iznošenja iskaza *Možeš li zatvoriti prozor?* slušatelj prepoznaje (doslovno) značenje rečenice od koje se sastoji (*Imaš li mogućnost zatvoriti prozor?*). Tek u sljedećoj fazi, nakon obrade doslovnoga značenja iskaza, slušatelj izvodi zaključak o smislu govornoga čina (*Zatvori prozor*). Računalna implementacija ovih procesa usmjerena je na upotrebu logike vjerovanja (eng. *belief logic*) za modeliranje ovakvoga inferencijskoga lanca.

Pristup utemeljen na vjerojatnosti (eng. *cue-based* ili *probabilistic approach*) predložili su Jurafsky i Martin (2000) inspirirani ulogom tzv. miga (eng. *cue*) u psiholingvističkim modelima usvajanja jezika i obrade rečenica. U ovim se modelima formalna obilježja rečenica od kojih se sastoje govorni činovi tumače kao *migovi* govornikovih komunikacijskih namjera. Kao i kod logički utemeljenih pristupa, odgonetanje govornikovih komunikacijskih namjera i u ovome pristupu uključuje inferenciju, ali ne i lanac koji polazi od (doslovnoga) značenja rečenice.

BDI-model (eng. *Belief, Desire, Intention Model*) jedan je od primjera modela pristupa utemeljenih na logici. Ovaj model temelji se na vjerovanjima, željama i namjerama sudionika komunikacije, a još se naziva i modelom utemeljenom na planu (eng. *plan-based model*). Zasniva se na pretpostavki da govornik u komunikaciji polazi od vjerovanjā o zajedničkome znanju koje dijeli sa svojim sugovornikom te ima mogućnost ažurirati svoja vjerovanja o sugovornikovim namjerama i vjerovanjima tijekom komunikacije.

Pogledajmo kako bi mogao izgledati inferencijski lanac interpretacije iskaza *Možeš li zatvoriti prozor?*, u kojemu slušatelj zaključuje da iskaz nije pitanje već implicitni zahtjev. U Primjeru 38 prikazan je mogući hodogram procesa zaključivanja o smislu govornoga čina (prema Jurafsky 2004).

Primjer 38

(1) X me pitao mogu li zatvoriti prozor.
(2) Pretpostavljam da je X kooperativan u komunikaciji (u smislu poštovanja Griceova kooperativnoga principa i načelā vođenja razgovara) te da stoga iskaz ima neku svrhu.
(3) X zna da mogu zatvoriti prozor te ne postoji razlog zbog kojega bi X doveo u pitanje moju sposobnost zatvaranja prozora.
(4) Stoga X-ova izjava vjerojatno ima neku neočitu ilokucijsku svrhu. Koja bi to svrha mogla biti?
(5) Početni uvjet za izvršavanje direktivnoga govornoga čina jest slušateljeva mogućnost izvršavanja govornikova zahtjeva.
(6) Stoga je X postavio pitanje o mojoj spremnosti za djelovanjem.
(7) Nadalje, X i ja smo u razgovornoj situaciji u kojoj je zatvaranje prozora uobičajeno i očekivano djelovanje.
(8) Stoga, u pomanjkanju nekih drugih mogućih ilokucijskih činova, X vjerojatno od mene zahtijeva da zatvorim prozor.

U okviru ovoga modela provodi se formalizacija vjerovanja, željenja, djelovanja i planiranja primjenom logičkih formalnih sustava. BDI-model može poslužiti i kao objasnidbeni model kojim se tumači zbog čega ljudi donose određene zaključke, a ne neke druge.

Iako BDI-model uključuje bogate strukture znanja i snažne tehnike planiranja, on ima i svoje nedostatke. Prije svega, utemeljen je na analizama pisanoga jezika te se zasniva na pretpostavci da svaki iskaz ima doslovno značenje (posredstvom kojega se dolazi do njegova smisla) – međutim, otvoreno je pitanje imaju li svi iskazi doslovno značenje. Osim toga, pojedini psiholingvistički eksperimenti ukazuju na to da ljudi paralelno obrađuju eksplicitno i implicitno iskazana značenja (Swinney i Cutler 1979 prema Jurafsky 2004).

Alternativni pristup tumačenju govornih činova sustav je temeljen na *migovima* (eng. *cue-based system*), koji se zasniva na pretpostavci da slušatelj koristi različite *migove* u ulaznim podacima pri interpretiranju govornih činova. U ovome modelu utemeljenome na vjerojatnosti površinski oblik ulaznih podataka (formalna obilježja iskaza) algoritmu nudi *migove* za izgradnju strukture. Vjerojatnosti su povezane sa značajkama površinskih oblika i nekim govornim ili dijaloškim činom. Iako i ovaj model u određenim fazama koristi inferencijske procese, algoritam donosi odluke i na temelju podataka, a ne isključivo pravila. Algoritmi za detekciju značajki promatranoga govornoga čina mogu se temeljiti na različitim izvorima znanja, tj. *migovima* (npr. prozodijski, leksički, kolokacijski, sintaktički ili diskursni *migovi*). Prednost ovoga modela u tome je što se većinom temelji na analizama govorenoga jezika.

Modeli utemeljeni na vjerojatnosti najčešće koriste algoritme nadziranoga strojnoga učenja.

> Kod nadziranog strojnog učenja algoritmu je dan skup označenih podataka, odnosno skup za učenje (*training set*), na kojem algoritam uči te vrši predviđanja na prethodno neviđenim podacima. (Bago 2014a: 156)

Skup za učenje predstavljaju ručno označeni govorni ili dijaloški činovi svakoga iskaza, a algoritam je statistički klasifikator koji uči prepoznati tipičnu kombinaciju značajki za svaki definirani tip govornih činova te za svaki iskaz odlučuje koji se govorni čin njime najvjerojatnije izvršava. Najjednostavniji način izgradnje vjerojatnosnoga modela zasniva se na utvrđivanju koje se riječi i izrazi češće pojavljuju u jednome dijaloškome činu naspram drugih (tzv. pristup n-gram). Primjerice, jednostavni Markovljev model za svaku riječ pohranjuje podatak koja je vjerojatnost njezina pojavljivanja ovisno o jednoj ili više određenih prethodnih riječi. Ulazni iskaz sastoji se od niza riječi *R*, a sustav odlučuje kojemu dijaloškome činu *d* pripada izračunavajući najveću vjerojatnost za dani niz riječi *R*. Uvjetna vjerojatnost *P(d|R)* izračunava vjerojatnost dijaloškoga čina *d* za niz riječi *R*. Prema Bayesovu teoremu uvjetna vjerojatnost može se raspisati kao u Primjeru 39:

Primjer 39

$P(d|R) = P(d)P(R|d)$

Jednadžba iz Primjera 39 izračunava vjerojatnost dijaloškoga čina umnoškom dviju vjerojatnosti čije se vrijednosti mogu dobiti iz skupa za učenje: vjerojatnost *P(d)* vjerojatnost je dijaloškoga čina u skupu za učenje; vjerojatnost *P(R|d)* vjerojatnost je niza riječi za odabrani dijaloški čin prema vjerojatnostima iz skupa za učenje. Iako je pristup n-gram jedan od najjednostavnijih vjerojatnosnih modela, i dalje može prepoznati neke riječi ili sintagme kao markere primjerice reformulacije (npr. *misliš na*) ili molbi (npr. *molim te*).

Osim ovoga najjednostavnijega primjera primjene vjerojatnosnoga modela n-gram na leksičkoj razini, isti je model moguće primijeniti i na ostale jezične razine znanja. Za interpretaciju i generiranje govornih i dijaloških činova primjenjuju se i neki složeniji vjerojatnosni modeli, iako je ta praksa tek u povojima. Vjerojatnosni modeli pružaju uvide u odnose između jezičnih oblika i njihovih funkcija. Razvojem prikladno označenih korpusa dostatne veličine sve će se više primjenjivati složeniji vjerojatnosni modeli, ali i hibridni modeli koji će primjenjivati vjerojatnosne i nevjerojatnosne modele.

4.3. Zaključci

Računalna pragmatika mlado je interdisciplinarno područje koje sve više napreduje razvojem jezičnih tehnologija. Osnovni izazov ovoga područja nedostatak je pragmatički anotiranih korpusa koji bi omogućili korištenje složenijih računalnih modela za njihovu obradu. Računalna pragmatika u hrvatskoj je znanosti tek u povojima. Za potrebe istraživanja predstavljenoga u ovoj knjizi razvijen je prvi specijalizirani korpus direktivnih govornih činova hrvatskoga jezika. Nadamo se da će taj poduhvat pridonijeti razvoju računalnopragmatičkih i korpusnopragmatičkih istraživanja hrvatskoga i drugih južnoslavenskih jezika.

U ovoj cjelini knjige nastojali smo ukratko prikazati čime se bavi obrada prirodnog jezika (OPJ) te kakva je njezina primjena u lingvistici – posebice pragmatici. U središnjem je poglavlju predstavljena računalna pragmatika: prikazan je njezin razvoj, istraživačka područja, metode i ciljevi. Pritom je posebna pažnja posvećena dvama tematskim područjima računalne pragmatike – izradi i primjeni korpusa za potrebe pragmatičkih istraživanja te računalnopragmatičkim pristupima interpretaciji i generiranju govornih činova. Naredne cjeline knjige posvećene su korpusnopragmatičkim analizama pragmatičkih pojavnosti u hrvatskome i srpskome jeziku.