3. Korpusi

Korpusi, kao anotirane arhive ljudske pisane i usmene komunikacije, objektivan su i pouzdan izvor za raznolike jezične analize. Bennett (2010) opisuje korpus kao veliku, strukturiranu zbirku izvornih primjera jezika koja je pohranjena u elektroničkom obliku. Kennedy (2014) definira korpus kao empirijsku osnovu za testiranje principa lingvističkih teorija. Na *Hrvatskom jezičnom portalu* navodi se da je korpus cjelovita zbirka podataka, dokumenata, građe za neku disciplinu (npr. korpus riječi hrvatskog jezika; korpus srednjovjekovnih dokumenata), a lingvistički bi to bio skup svih pojavnica koji se izrađuje računalnim putem za uporabu u konkordancijama, tezarijima i sl. (npr. Hrvatski nacionalni korpus), što je u užem smislu računalni korpus.

Porijeklo prikupljanja jezičnih korpusa datira još iz 13. stoljeća, kada su istraživači ručno indeksirali riječi iz Biblije. Konkordancija je nastala iz praktične potrebe da se drugim proučavateljima Biblije olakša pretraga riječi po abecednom redu i odlomcima u kojima se nalaze. Cilj je bio dokazati da Biblija nije tekst iz više izvora, već skladna božanska poruka (McCarthy i O'Keeffe, 2010). Kroz povijest su predmetom konkordancija i indeksiranja najčešće bila književna djela, a unatrag sedamdesetak godina, s razvojem tehnologije, korpusi se šire na sva područja ljudske komunikacije, i tipične (poput dječjeg govora) i atipične (npr. govora osoba s afazijom), te tvore opće ili specijalizirane korpuse unutar neke države ili jezične zajednice, iako nisu rijetki ni višejezični korpusi (npr. Babel, koji sadrži zapise iz 22 jezika (Gales i sur., 2014)). Zanimljivo je da se prvi elektronički korpus također veže uz proučavatelje kršćanske literature, naime isusovački svećenik Robert Busa izradio je elektroničko lematizirano kazalo cjelovitih djela svetoga Tome Akvinskog krajem 70-ih godina 20. stoljeća (McCarthy i O'Keeffe, 2010).

Iako su djela ranih proučavatelja Biblije i književnosti (npr. Shakespeareova djela) bila preteča pretraživanja riječi i indeksiranja, američki predstrukturalisti preteča su korpusa, ne samo u smislu prikupljanja podataka već stoga što stavljaju jezične podatke u srž lingvističkih istraživanja (Leech, 1992 prema McCarthy i O'Keeffe, 2010). Korpusna je lingvistika svoj pravi potencijal doživjela razvojem tehnologije, posebice interneta jer su se količina jezičnih podataka i način njihova

prikupljanja drastično promijenili i povećali (vidi npr. Cambridge International Corpus).

Danas se korpusi ne koriste samo za izučavanje književnih djela već imaju različite uporabe i koriste se u različitim područjima: za podučavanje jezika i učenje (Campbell i sur., 2007), u prevođenju, forenzičnoj lingvistici, sociolingvistici, pragmatici, suvremenoj dijalektologiji, dijagnostici (Haulcy i Glass, 2021) itd.

Osim za potrebe jezičnih istraživanja, govorni korpusi važni su i zbog očuvanja govora, posebice kada je riječ o manjim jezicima i govorima koji se oslanjaju više na usmenu književnost te nisu zastupljeni u medijima pa postoji manje zapisa na temelju kojih bi se njihovi govori mogli čuvati i istraživati. U tradicionalnim dijalektološkim istraživanjima u Hrvatskoj vrlo često nedostaju čak i sociodemografski podatci o govornicima (pa čak i njihov broj), a zvučni zapisi dijalektoloških govora također nisu dostupni (Biočina, 2019). Potonje umanjuje eksperimentalnu ponovljivost tih istraživanja, ali i onemogućava diseminaciju znanja. Tek neka novija dijalektološka istraživanja nude podatke o govornicima (npr. Bašić, 2012; Galović, 2017; Biočina, 2019), međutim, tek rijetki autori daju uvid u snimke govora i tako onemogućavaju daljnja istraživanja dijalekata, ali i njihovo očuvanje. Dostupnost za šire istraživačke potrebe jedna je od ključnih prednosti javno dostupnih korpusa jer omogućavaju diseminaciju znanja, uštedu vremena i iscrpniju jezičnu analizu prikupljenog materijala.

Opći korpusi trebali bi sadržavati što više stilova i registara pojedinog jezika, uključujući i uzorke govorenog jezika (Campbell i sur., 2007). Stoga možemo reći da bi trebali dobro predstavljati jezik u cjelini, odnosno biti reprezentativni. Korpusi mogu nastati zbog mnogo različitih razloga, a razlozi pomažu odrediti veličinu korpusa, stil i sadržaj. Reprezentativnost korpusa i njegova veličina, ključni su jer određuju koja se istraživačka pitanja mogu postaviti (Lavid, 2010). Međutim, iako je reprezentativnost upravo ono što izdvaja korpuse od ostalih vrsta prikupljanja jezičnih podataka, koncept reprezentativnosti iznenađujuće je neprecizno definiran u korpusnoj lingvistici (Pastor i Seghiri Domínguez, 2010) te varira od područja do područja i s obzirom na vrstu korpusa i dizajn (jezičnog) korpusa (Davis, 2018). Govorni korpusi, primjerice, skloniji su uključivanju neprofesionalnih govornika, stoga postoji velika diskrepancija u veličini između korpusa pisanog i govorenog jezika (Kuvač Kraljević, Hržica, Kolografić Belić, 2020: 127). Pritom su korpusi pisanog jezika uglavnom veći te često sadrže novinske članke, književna djela i sl., čiji su autori školovani pisci ili im je pisanje profesija. Potonje umanjuje izvornost autohtonoga jezičnog ostvaraja, ali omogućuje brže prikupljanje veće količine materijala. Kako bi osigurali reprezentativnost, opći korpusi teže uzorkovanju velikog raspona žanrova. Primjerice, govoreni dio korpusa CEC (Cambridge English Corpus) sadrži uzorke govora iz svakodnevnog razgovora, radijskih emisija i TV programa, prezentacija, javnih nastupa, sastanaka i predavanja. Jedan je od općih

korpusa hrvatskoga jezika primjerice Hrvatski nacionalni korpus – hrvatski korpus pisanog jezika koji se sastavljao u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu (Tadić, 1998; 2009).

Za razliku od općeg korpusa, specijalizirani korpus može težiti skupljanju demografski šarolikog spektra govornika, uključujući različite socioekonomske skupine, geografsku lokaciju i jezični status (Kuvač Kraljević i Hržica, 2016; Kuvač Kraljević, Hržica, Kologranić Belić, 2020; Varošanec-Škarić, Bašić i Biočina, 2022). Primjerice TalkBank, osim što sadrži korpuse različitih jezika, okuplja i nekoliko baza kliničkih korpusa, kao što je AphasiaBank (MacWhinney i sur., 2011). TalkBank sadrži i nekoliko specijaliziranih korpusa iz Hrvatske koje prikupljaju Kuvač Kraljević i suradnici:

- Hrvatski korpus dječjeg jezika (HKDJ, Kovačević, 2002)
- Hrvatski korpus govornog jezika (HrAL, Kuvač Kraljević i Hržica, 2016)
- Hrvatski diskursni korpus govornika s afazijom (CroDA, Kuvač Kraljević, Hržica i Lice, 2017)
- Hrvatski korpus neprofesionalnog pisanog jezika (HKNPJ, Kuvač Kraljević, Hržica, Kologranić Belić, 2020).

Hrvatski korpus govornog jezika (HrAL) sadrži uzorke spontanog govora 617 govornika iz svih hrvatskih županija prikupljane od 2010. do 2016. godine (Kuvač Kraljević i Hržica, 2016). U TalkBanku se može naći pod *Conversational analyses corpora* unutar pododjeljka *Conversational Bank* (https://ca.talkbank.org/). Taj korpus neprofesionalnoga pisanog hrvatskog jezika sadrži diskurs tipičnih govornika i govornika s jezičnim poremećajima (Kuvač Kraljević, Hržica, Kologranić Belić, 2020), čime predstavlja jedinstvenu bazu pisanog jezika jer uzorkuje jezik koji su producirali neprofesionalni govornici iz različitih krajeva Hrvatske s tipičnim i netipičnim jezičnim ponašanjem prikupljen u identičnim uvjetima.

Od 2022. godine za široku uporabu dostupna je i zvučna baza Bazvuka – višejezični korpus zvučnih govornih zapisa visoke kvalitete. Autorice su Gordana Varošanec-Škarić, Iva Bašić i Zdravka Biočina. Korpus je prikupljen u razdoblju od 2011. do kraja 2020. godine i sadrži govorne zapise 621 izvornog govornika hrvatskog, slovenskog, bosanskog i srpskog govora. Sva su snimanja provedena visokokvalitetnom opremom u studijskim uvjetima ili u sobama sa sniženom razinom buke. Govorni materijal sastoji se od čitanja kraćih tekstova (frikativni i nefrikativni tekst), forme poluspontanog govora (vođeni razgovor) i/ili čitanja niza riječi i rečenica. Uvjeti snimanja bili su slični (visokokvalitetna oprema, prostorije bez buke, udaljenost usta od mikrofona – 10 cm, pod kutom od 45 °) i provodili su ih profesionalci, fonetičari, što je bilo važno zbog poštivanja navedenih uvjeta snimanja. Zvučni zapisi montirani su u programu za akustičku obradu zvuka Praat (Boersma i Weenink, 2020). Iako su se snimanja provodila u kontroliranim uvjeti-

ma, određene govornike nije se moglo uključiti u bazu zbog dislalije, nekvalitetne snimke, izvanjske buke, uzlazne intonacije i sl. Neki govornici nisu bili dovoljno razgovorljivi pa nije prikupljeno dovoljno spontanog govora ili nisu htjeli snimiti dio materijala. Konkretnije, nekoliko starijih govornika nije htjelo čitati tekst, vjerojatno zbog slabije vještine čitanja ili slabijeg vida.

Sveukupno korpus čini 45,25 % govornika i 54,75 % govornica, a ukupno trajanje govora iznosi 84,18 sati (prosječno trajanje po govorniku iznosi 6,03 minute). Korpus hrvatskog govora prikupljen je u 13 gradova diljem Hrvatske te na otoku Braču (10 mjesta). Govornici srodnih jezika snimljeni su u glavnim gradovima zemalja (Beograd, Ljubljana, Sarajevo). Baza je oblikovana tijekom 2021. i početka 2022. godine (https://bazvuka.ffzg.unizg.hr/).

U okviru suradnje Odsjeka za fonetiku i KBC-a Sestre Milosrdnice (2017. –) te dvaju institucijskih istraživačkih projekata voditeljice prof. dr. sc. Gordane Varošanec-Škarić (Odsjek za fonetiku, Filozofski fakultet Sveučilišta u Zagrebu) u razdoblju od 2022. do 2023. godine prikupljen je korpus od 50 atipičnih glasova, koji se i dalje nadograđuje u okviru institucijskih istraživačkih projekata voditeljice doc. dr. sc. Ive Bašić (2024. i 2025. g.). Svi su govornici snimljeni specijaliziranom opremom za akustička snimanja u Studiju za akustička snimanja pri Odsjeku za fonetiku. Od govornih materijala bile su zastupljene fonacije vokala [a], [i] i [u] u trostrukim izvedbama (a ponekad i manje, ovisno o tome koliko je govornik/pacijent mogao sudjelovati s obzirom na razinu i vrstu oštećenja glasa), čitaći govor (tekst *Sjeverni ledeni vjetar i sunce*, nefrikativni tekst, specijalizirani tekst za mjeru kepstralnoga vokalnog vrhunca te kvazispontani govor na temelju intervjua govornika i voditelja snimanja/fonetičara). Većina je govornika (pacijenata) snimljena preoperativno, a manji je udio snimljen i postoperativno.

Kao što je prethodno navedeno, korisno je, zbog širokog spektra fonetskih analiza, imati i korpus čitaćeg govora (tekst, liste riječi) kao i korpus spontanog govora (najčešće kvazispontani govor koji dobivamo intervjuom s govornikom (dijalog, ali može biti i monolog)). Prednost je intervjua u govornikovu životnom okruženju što dobivamo izvorniji govor, dok u studiju za snimanja i tihim sobama dobivamo bolji akustički signal odnosno kvalitetniju snimku.

Raznolikost materijala koji se prikuplja ovisi o tipu i cilju korpusa. Primjerice, korpusi koji se koriste za forenzične potrebe često su prikupljali snimke telefonske transmisije (danas zbog 5G mreže ne dolazi do distorzije vrijednosti pa nema potrebe za takvim snimkama), prikrivenog govora, govora s maskom i sl. AHUMADA (Ortega-Garcia, Gonzalez-Rodriguez, Marrero-Aguiar, 2000) je upravo jedan takav korpus španjolskog jezika koji sadrži čitanje tekstova različitom brzinom govora, spontani govor, govor s različitim mikrofonima i preko telefona, snimke istih govornika snimljene u šest različitih snimanja, dijalektalne varijacije govornika itd. Zanimljiv je i korpus UT-Scope (Ikeno, Varadarajan i Hansen, 2007), koji sadr-

ži govorne zapise u bučnom i tihom okruženju 30 govornika engleskog jezika iz SAD-a. Lombardijev efekt izazvan je reprodukcijom buke kroz slušalice, što osigurava kvalitetu same snimke, a postiže željeni učinak govora u buci kod govornika. Zatim, dva korpusa sa snimkama stvarnih slučajeva, jedan iz Njemačke (Solewicz i sur., 2012) te korpus NFI-FRITS, koji sadrži podatke iz stvarnih telefonskih presretanja (van der Vloed i sur., 2014). Korpus SweEval2016 noviji je manji testni set prikladan za forenzičnu usporednu procjenu govornika koji se sastoji od 23 snimke govora na mobilnim telefonima (10 poznatih i 13 nepoznatih govornika) (Lindh, Åkesson i Sundqvist, 2016). Korpusi koji se koriste u forenzične svrhe mogu primjerice sadržavati i samo govornike jednog spola, najčešće muškog (npr. The DyViS database Nolana i suradnika (2009)).

Iako nam se iz ovoga kratkog pregleda specijaliziranih forenzičnih korpusa može činiti da postoji velik dijapazon takvih korpusa, forenzični fonetičari iz Velike Britanije upozoravaju kako nedostaje forenzičnih korpusa sa stvarnim govornim zapisima (Hughes, 2014; Ross, French i Foulkes, 2016). Korpuse govorenog jezika neprofesionalnih govornika znatno je složenije prikupiti nego korpuse pisanog jezika jer zahtijevaju obučene profesionalce, kvalitetnu opremu i studija za snimanje ili tihe sobe (zbog kvalitete snimki). Taj problem pokušava se riješiti modernom tehnologijom, primjerice uporabom robota AIBO u prikupljanju korpusa dječjeg govora (Batliner i sur., 2004). Govorni korpusi koji se prikupljaju u forenzične svrhe najčešće nisu transkribirani, odnosni nisu lingvistički kodirani, stoga se u okvirima korpusne lingvistike ne bi smatrali korpusima u klasičnom smislu riječi nego o zbirkama govora.

U Hrvatskoj je nedostatak korpusa govorenog jezika puno očitiji, iako je Bazvuka umanjila diskrepanciju u veličini između korpusa pisanoga jezika i zvučnih baza. Savjeti za snimanje govornika i montažu zvuka predstavljeni su u prethodnom poglavlju. Uvijek je poželjno da snimanje provodi sam istraživač-fonetičar ili općenito fonetičar ili osoba obučena za snimanje. Snimanje se za sve govornike treba provoditi u približno jednakim uvjetima i s istom opremom. Sociodemografskim upitnikom uputno je prikupiti što više informacija o govorniku koje onda mogu biti korisne u formiranju baze govora i njezine podjele po raznim izvanjezičnim obilježjima govornika.

Važnost korpusa za istraživanja govora neizmjerna je, osim što omogućuje očuvanje govora i detaljnije analize istoga govornog materijala, omogućuje i izračun akustičkih mjera na relevantnom uzorku te predstavlja vrijedan izvor podataka i za fonetiku (npr. podatci za populaciju) i forenzičnu fonetiku, ali i za umjetnu inteligenciju (npr. za učenje programa za automatsko prepoznavanje govora, v. npr. Common voice (Ardila i sur., 2020)).