

Gordana Varošanec-Škarić, Iva Bašić

Odsjek za fonetiku

Filozofski fakultet Sveučilišta u Zagrebu, Hrvatska

gvarosan@ffzg.unizg.hr, ibasic@ffzg.unizg.hr

„Bazvuka” – zvučna baza govora izvornih govornika hrvatskoga jezika i srodnih jezika

U razdoblju od 2011. do kraja 2020. godine snimljen je 621 izvorni govornik hrvatskoga, srpskoga, slovenskoga i bosanskoga jezika. Govorni su stilovi obuhvatili čitaći govor (riječi, rečenice, kraće tekstove), spontani govor u formi intervjua i/ili fonacije vokala. Snimalo se u sličnim uvjetima (visokokvalitetnom opremom u prostorijama bez buke, udaljenost mikrofona bila je 10 cm od usta, pod kutom od 45°). Snimljeno je 45,25 % muškaraca i 54,75 % žena, a ukupno trajanje govora iznosi 84,18 sati (po govorniku oko 6,03 min). Korpus hrvatskoga govora prikupljen je u 13 gradova diljem Hrvatske te na otoku Braču u deset mjesta. Govornici srodnih jezika snimljeni su u glavnim gradovima zemalja (Beogradu, Ljubljani i Sarajevu). Govorne snimke u zvučnoj bazi *Bazvuka* montirane su u programima za akustičku obradu zvuka i prilagođene za ulazak u infrastrukturu (HR-)CLARIN, koja je sastavni dio europskih istraživačkih infrastruktura. *Bazvuka* je oblikovana tako da se longitudinalno može nadopunjavati dodatnim snimkama govora, a slobodan će pristup imati fonetičari, lingvisti, studenti te drugi znanstvenici i AAI@EduHr korisnici. U radu je prikazan primjer istraživanja provedenog na temelju korpusa govornih zapisa iz baze *Bazvuka*, u kojem je ispitana međujezična razlika u ritmicitetu između govornika hrvatskoga, srpskoga i bosanskoga jezika.

Ključne riječi: *Bazvuka*, zvučna baza govora, hrvatski jezik, srodni jezici

1. Uvod

1.1. Ciljevi za izgradnju zvučne baze „Bazvuka”

Uspostavljanje zvučne baze govora¹ pretpostavlja prethodna određenja korpusa govornika (jezik, broj govornika, kriteriji za izbor govornika), uvjeta snimanja (na temelju istraživačkoga iskustva određuje se koji će biti

¹ U engleskom jeziku uvriježeni su nazivi sa sličnim govornim materijalima ‘speech corpus’, ‘spoken corpus’ te ‘corpus of speech and language’.

uvjeti snimanja i nadzor za vrijeme snimanja kako bi se ujednačili uvjeti snimanja koji bi mogli utjecati na kvalitetu snimke), trajanja snimanja (radi pragmatičke provedivosti prikupljanja govornoga materijala) i na kraju pretpostavlja izgradnju baze (grafički dizajn, odabir repozitorija te izradu sučelja). U potonjem smislu važno je prethodno odrediti u kojem će se obliku unositi audiosnimke, kako će se montirati i oblikovati sa svrhom bolje iskoristivosti i dostupnosti svim istraživačima u otvorenom pristupu, a ne samo projektnim istraživačima.

Zvučna baza govora definirana je kao organizirani skup audiopodataka pohranjenih i dostupnih na elektronički način. Do sada je montirano i pohranjeno 1019 govornih zapisa izvornih govornika hrvatskoga jezika i srodnih jezika (srpskoga, slovenskoga, bosanskoga). Prioritetna je svrha izgradnje i ustanovljenja zvučne baze dostupnost uređenih govornih materijala i široj znanstvenoj te stručnoj zajednici. Dakle, bazom će se u istraživačke svrhe moći koristiti fonetičari, lingvisti, studenti (npr. za potrebe seminarskih, diplomskih, doktorskih radova), ali i drugi AAI@EduHr korisnici. U inozemnim su bazama s govornim materijalom ponekad dostupni i transkripti, pa je stoga jedan od budućih ciljeva baze i transkribiranje spontanoga govora, što može biti korisno u istraživačkoj sinergiji lingvista i fonetičara.

1.1.1. Testiranje i oblikovanje zvučne baze govora „Bazvuka”

Završnom osmišljavanju baze prethodio je postupak testiranja pomoću testnoga korisničkog računa na *Dataverse* sustavu, koji se pokazao vrlo fleksibilnim i konfigurabilnim za uporabu te ima vrlo dobru podršku za programski pristup (API). Prednosti su navedenoga sustava i široko prihvaćeni identifikatori koji se dodjeljuju objektima (DOI) te da je sve što je potrebno sadržano na istome mjestu, pa na taj način i glavni istraživač i članovi projekta ne ovise o raspoloživosti razvojnih inženjera na SRCE-u (što bi znatno produljivalo rok u kojem bi baza mogla biti otvorena i drugim istraživačima/korisnicima). Za tu fazu trebalo je na ekonomičan i korisnicima jednostavan način oblikovati način unosa i pregled metapodataka koji su korisni svim istraživačima. Ova će zvučna baza doprinijeti i tome da naša znanstvena zajednica (fonetičara, lingvista i filologa) može koristiti bazu i korpus snimljenih govornih zapisa, a bit će omogućen pristup i za europske istraživače, primjerice pri usporedbi spontanoga govora na hr-

vatskom jeziku i nekom drugom jeziku. S obzirom na raznovrsnost zastupljenih govornih stilova, korpus govornih zapisa bit će pogodan za različita istraživanja. Dio korpusa u bazi koji obuhvaća izvorne hrvatske govornike longitudinalno će se nadopunjavati govornim zapisima sa svrhom da bude resurs i u idućem razdoblju za istraživanja govora na segmentnoj i supra-segmentnoj razini.

Osim planiranoga longitudinalnog proširivanja baze govora u budućim istraživačkim razdobljima, osiguran je i ulazak u istraživačku infrastrukturu (HR-) CLARIN² u skoroj budućnosti. Istraživačka struktura ima različita značenja, a za humanističke i društvene znanosti esencijalno je da to bude mreža arhiva u kojima su pohranjeni primarni istraživački podatci i sredstva za njihov pristup. Projekt CLARIN započeo je 2008. g. u okviru Paneuropske inicijative u projektu 7. okvirnog programa Europske istraživačke infrastrukture za humanističke i društvene znanosti (*Common Language Resources and Technology Infrastructure*) u kojem je od 2012. sudjelovao i hrvatski lingvist Marko Tadić (<https://www.clarin.eu/>).

U svrhu forenzičnih fonetskih mjerenja nova zvučna baza *Bazvuka* bit će korisna i u usporedbi govora pri prepoznavanju govornika – ciljanih govornika (osumnjčenika) u odnosu na govornike iz govornoga područja ciljanih govornika. Govornim snimkama osumnjčenika u zatvorenoj bazi moći će pristupiti samo fonetičari, članovi projekta. Na temelju gradova i mjesta moći će se oblikovati histogrami distribucije mjera fundamentalne frekvencije (F0) govornika, izračunati raspodjele rezultata za medijan F0, bazičnu i alternativnu fundamentalnu frekvenciju (*baseline* i *alternative baseline*), što je u forenzične svrhe učinjeno za referentne uzorke govornika (Varošanec-Škarić, 2019). Na spontanome će govoru biti moguće istražiti vrste disfluentnosti (prekidanja, preklapanja, punjače, artikulacijske pogreške, zamjene riječi, umetanja itd.) te ispitati jezičnu, dobnu i spolnu varijabilnost među govornicima istoga, ali i različitih jezika.

2 Istraživačka (HR-)CLARIN infrastruktura sastavni je dio europskih istraživačkih infrastruktura.

1.1.2. Očekivani doprinos

Očekuje se da će znanstveni doprinos ove zvučne baze biti u razvijanju i izgradnji novoga korpusa zvučne baze hrvatskoga jezika i srodnih jezika (srpski, slovenski, bosanski) koji može biti koristan pri sociolingvističkim i fonetskim forenzičnim istraživanjima. Očekuje se da će zvučna baza govora biti korisna i pri analizi govora u odnosu na varijable varijeteta (gradski i mjesni varijeteti), spola i roda te pri unaprjeđenju procedure i resursa za forenzična istraživanja našim i europskim istraživačima.

1.2. Prethodna istraživanja

Prethodna istraživanja mogu se podijeliti u prvo trogodišnje projektno razdoblje na projektu pod nazivom *Forenzična fonetika: slušno prepoznavanje i zvučna analiza glasova* (2007. – 2013., glavne istraživačice Gordane Varošaneć-Škarić) te iduće devetogodišnje razdoblje obuhvaćeno u devet odobrenih financijskih potpora istraživanjima Sveučilišta u Zagrebu, redom: *Forenzička fonetika: standardiziranje zvučnih postupaka* (2013), *Forenzička fonetika 2: standardiziranje zvučnih postupaka* (2014), *Forenzička fonetika 3: Mjere fundamentalne frekvencije (f0) muških govornika* (2015), *Forenzička fonetika 4: Mjere fundamentalne frekvencije (f0) hrvatskih i srpskih govornika* (2016), *Forenzička fonetika 5: Usporedba akustičkih parametara općega hrvatskoga izgovora s dijalektalnim izgovorom i drugim jezicima* (2017), *Forenzična fonetika – međujezični i idiosinkratični prozodijski parametri* (2018), *Punjači i disfluentnosti u surječju forenzične fonetike* (2019), *Vrste disfluentnosti i strategija prekidanja u javnome govoru* (2020), *Zvučna baza govora izvornih govornika hrvatskoga jezika i srodnih jezika* (2021).

U projektnim okvirima u razdoblju od 2012. do 2019. g. izrađena su i obranjena tri doktorska rada istraživačica koje su provodile dijelove terenskih snimanja s fokusom na profiliranju i prepoznavanju govornika, a sva su tri rada uključivala i zvučne analize: radovi Gabrijele Kišiček (2012), Ive Bašić (2018) i Zdravke Biočine (2019). Tijekom prikupljanja korpusa (od 2009. g. dalje) objavljen je niz znanstvenih radova članica projekta, primjerice s fokusom na indeksnim osobinama govornika nekih hrvatskih gradova (Varošaneć-Škarić i Kišiček, 2009), na usporedbi slušne procjene i zvučnih vrijednosti F0, formantnih frekvencija (F1–F4) s regionalnom procjenom hrvatskih govornika (Kišiček i Varošaneć-Škarić, 2011), na utjecaju

kvalitete glasa na procjenu regionalnosti, društvenih i psiholoških osobina govornika (Kišiček i Varošaneć-Škarić, 2016). Važan je fokus bio na procjeni lingvističke bliskosti i akustičkih osobina hrvatskoga i srpskoga jezika (Varošaneć-Škarić, 2017), potom u forenzičnom surječju na statističko-akustičkom istraživanju indeksa sličnosti i različitosti govornika hrvatskoga jezika u različitim zvučnim uvjetima (Varošaneć-Škarić i sur., 2014), zvučne osobine hrvatskih kardinalnih vokalskih formanata (F1, F2 i F3) (Varošaneć-Škarić i Bašić, 2015) i razlike u formantima i fundamentalnoj frekvenciji prema spolu u usporedbi govornika hrvatskoga i srpskoga jezika (Bašić i Varošaneć-Škarić, 2019) te su testirane razlike u akustičkim parametrima u različitim uvjetima snimanja (Bašić i Biočina, 2020). U surječju prepoznavanja govornika treba spomenuti i prepoznavanje dijalektalnoga govora čak deset mjesta otoka Braća (Biočina i Varošaneć-Škarić, 2017) i zvučno istraživanje prozodije pražničkoga govora (Biočina i sur., 2018). Brojna istraživanja u svrhu usporedbe skupina govornika u varijablama jezika, spola, dobi bila su važna za utvrđivanje govornika nužnih za određena zvučna istraživanja i za postizanje većega stupnja statističke probabilnosti. Također, nastojali su se osigurati svi nužni uvjeti za uspostavljanje baze nužne za zvučnu forenzičnu usporedbu među govornicima, navedenu u sveučilišnom udžbeniku iz područja forenzične fonetike (Varošaneć-Škarić, 2019).

2. Metodologija

2.1. Snimanje korpusa

Dio zvučnih zapisa snimljen je u studiju za akustička snimanja na Odsjeku za fonetiku Filozofskoga fakulteta Sveučilišta u Zagrebu visoko kvalitetnom opremom (AKG, model C 414 B/ULS: Large Diaphragm Condenser Microphone; frekvencija uzorkovanja: 44 100, 16 bita; frekvencijskoga raspona od 20 do 20 000 Hz), a u dijelu terenskih snimanja snimalo se u prostorijama sa sniženom razinom buke (uklonjena je buka računala i druge računalne opreme, klimatizacijskih uređaja itd.), u kojima su također korišteni visokokvalitetni snimači (Marantz, model PMD 660 i ZOOM, model H4n) te mikrofoni (Røde, model NT5_Compact Cardioid Condenser Microphone i AKG mikrofoni, model AKGC 391). Svi su govornici snimljeni na udaljenosti od 10-ak cm od mikrofona, pod kutom od 45°. Zvučni su se zapisi montirali

u programima za akustičku obradu zvuka Praat (Boersma i Weenink, 2010), Cool Edit Pro te Audacity.

2.2. Korpus

U svrhu prikupljanja korpusa zvučne baze, snimanje govornika u Zagrebu i terenska snimanja provedena su u razdoblju od 2010. do 2021. godine. Studijska su snimanja nastavljena i nakon toga, a do sada su u bazu postavljene snimke govora govornika koji su prikupljeni do kraja 2018. Ukupno je u cijelome korpusu prikupljen govor 621 govornika (M = 281 (45,25 %), Ž = 340 (54,75 %)), čije ukupno trajanje govornih zapisa iznosi 84,18 sati (prosječno trajanje snimke po govorniku – 6,03 min), dok ukupna količina podataka iznosi 27 569,05 MB (prosječno 39,51 MB po govorniku). Korpus hrvatskoga govora prikupljan je u većim urbanim središtima, uglavnom u razdoblju od 2010. – 2017. U Zagrebu je snimljen 151 govornik (M = 92 (60,93 %), Ž = 59 (39,07 %)), čije ukupno trajanje iznosi 19,9 sati (prosječno trajanje snimke od 9 min po govorniku), dok ukupna količina podataka iznosi 5682 MB (prosječno 42,94 MB po govorniku). U Varaždinu je snimljeno 10 govornika (M = 6, Ž = 4), u Čakovcu 24 govornika (M = 6, Ž = 18), u Splitu 18 govornika (M = 5, Ž = 13), u Rijeci 15 govornika (M = 6, Ž = 9), u Zadru 15 govornika (M = 6, Ž = 9), u Šibeniku 22 govornika (M = 7, Ž = 15), u Dubrovniku 19 govornika (M = 5, Ž = 14), u Osijeku 40 govornika, podjednako muškaraca i žena, u Poreču 1 govornik (Ž = 1), u Opatiji 2 govornika (M = 2), u Puli 27 govornika (M = 6, Ž = 21), u Slavonskome Brodu 5 govornika (M = 2, Ž = 3) i u Vukovaru 5 govornika (Ž = 5). U svrhu istraživanja prepoznavanja govornika otoka Brača ukupno je snimljeno 150 govornika dijalektalnoga izgovora na otoku Braču (M = 69 (46 %), Ž = 81 (54 %)), čije ukupno trajanje govornih zapisa iznosi 24,60 sati (prosječno trajanje snimke po govorniku – 9,84 min), dok ukupna količina podataka iznosi 7708 MB (prosječno 51,38 MB po govorniku).

Od srodnih su jezika prikupljeni zvučni zapisi izvornih govornika srpskoga, slovenskoga i bosanskoga govora. Muški i ženski govornici hrvatskoga i srpskoga jezika čiji je govor procijenjen neregionalnim (alokalnim) snimljeni su od 2014. do kraja 2016. u Zagrebu. Verifikator za hrvatski jezik bila je prof. dr. sc. Gordana Varošaneć-Škarić i u dijelu korpusa dr. sc. Iva Bašić s Odsjeka za fonetiku Filozofskoga fakulteta u Zagrebu, a verifikator za srpski

jezik bio je prof. dr. sc. Dušan Marinković s Odsjeka za slavistiku Filozofskoga fakulteta u Zagrebu. Na Braču su 2017. godine snimljeni govornici koji su predstavljali mjesne govore, a iste godine u Ljubljani snimljeni su govornici neregionalnoga izgovora za slovenski jezik.

Verifikator je za slovenski jezik bio prof. dr. sc. Hotimir Tivadar s Odsjeka za slovenistiku (Filozofska fakulteta, Univerza v Ljubljani, Oddelek za slovenistiko).

U Sarajevu je 2018. snimljeno 35 govornika (za hrvatski, bosanski i srpski jezik; verifikator je bila prof. dr. Amela Šehović, voditeljica Centra za bosanski, hrvatski i srpski jezik na Filozofskom fakultetu Univerziteta u Sarajevu). Korpus govornika srpskoga jezika koji su snimljeni u Beogradu čine 82 govornika (M = 46 (56,10 %), Ž = 36 (43,90 %)), čije ukupno trajanje govornih zapisa iznosi 15,55 sati (prosječno trajanje snimke po govorniku – 11,39 min), dok ukupna količina podataka iznosi 4630 MB (prosječno 56,25 MB po govorniku), zatim 40 govornika iz Ljubljane (M = 19 (47,50 %), Ž = 21 (52,50 %)), čije ukupno trajanje govornih zapisa iznosi 6,56 sati (prosječno trajanje snimke po govorniku – 9,90 min), dok ukupna količina podataka iznosi 1913 MB (prosječno 47,75 MB po govorniku) te 35 govornika iz Sarajeva (M = 3 (8,57 %), Ž = 32 (91,43 %)), čije ukupno trajanje govornih zapisa iznosi 4,19 sati (prosječno trajanje snimke po govorniku – 7,30 min), dok ukupna količina podataka iznosi 2612 MB (prosječno 75,10 MB po govorniku).

S obzirom na to da je u odnosu na korpus najviše govornika hrvatskoga jezika (75 %), pretpostavlja se da će se baza najviše koristiti za istraživanja govora hrvatskih govornika. S druge strane, s obzirom na to da je prikupljeno dovoljno ispitanika za istraživanje nekih prozodijskih i akustičkih parametara i za druge jezike (pogotovo za srpski, s ukupnim udjelom od 13 %), baza će biti korisna svim istraživačima kojima su u fokusu kontrastivne jezične analize.

2.2.1. Sociodemografski upitnik

Uspostava svake baze koja iznosi podatke o ispitanicima, pa tako i zvučne baze govora koja će biti otvorenoga pristupa, pretpostavlja osiguravanje etičnosti pri snimanju, ali, naravno, i nakon snimanja. Identitet je svih govornika zaštićen i anoniman. Svi su sudionici snimanja bili upoznati s

ciljem prikupljanja korpusa govornoga materijala i obaviješteni su o vrsti govornoga materijala koji se prikuplja. Nakon informiranja i uvida u upitnike, svaki je sudionik ispunio sociodemografski upitnik i potpisao pristanak na sudjelovanje u istraživanju. Sociodemografski upitnik sadrži opće demografske podatke (o spolu, rodu, godini rođenja, dobi, mjestu rođenja, mjestu prebivališta, stručnoj spremi, konzumaciji alkoholnih pića i duhanskih proizvoda), navodi naziv istraživanja, sadrži podatke o glavnom istraživaču i drugim članovima istraživanja, svrsi istraživanja, njegovome tijeku i trajanju. Upitnik također sadržava i informaciju o obvezi istraživača da će identitet govornika/sudionika istraživanja biti zaštićen. Osim toga, u upitniku je naznačeno da u bilo kojem trenutku govornik može odustati od snimanja te da nakon snimanja može zatražiti i dobiti audiosnimke.

Snimanje govornika i unos govornoga materijala u bazu provodili su se sukladno navedenim etičkim standardima za provedbu istraživanja. Identitet govornika bit će poznat samo voditelju istraživanja i istraživačima intervjuistima. Faze projektnoga istraživanja odobrilo je Ministarstvo znanosti i obrazovanja Republike Hrvatske ili sveučilišna tijela u okviru brojnih istraživačkih financijskih potpora. Etika istraživanja provodit će se i u nastavku studijskoga i terenskoga prikupljanja govora izvornih govornika hrvatskoga jezika.

2.2.2. Govorni materijal

Za potrebe korpusa snimljen je raznovrstan govorni materijal da bi se zadovoljio prioritetni cilj stvaranja zvučne baze te ciljanih ispitivanja, primjere zasebnih korpusa za kontrastivne analize i prepoznavanje govornika. Tako su, uz čitanje tekstova (tzv. nefrikativnog i frikativnog; Škarić, 1993³), za potrebe mjerenja vokalskih formanata snimljene i niske riječi te rečenica s logičkim naglašavanjem ciljane riječi. Spontani govor ispitanika sniman je na temelju intervjua s fonetičarom. Intervjuist je postavljao pitanja koja su uobičajena u sociofonetskim, sociolingvističkim i forenzičnim istraživa-

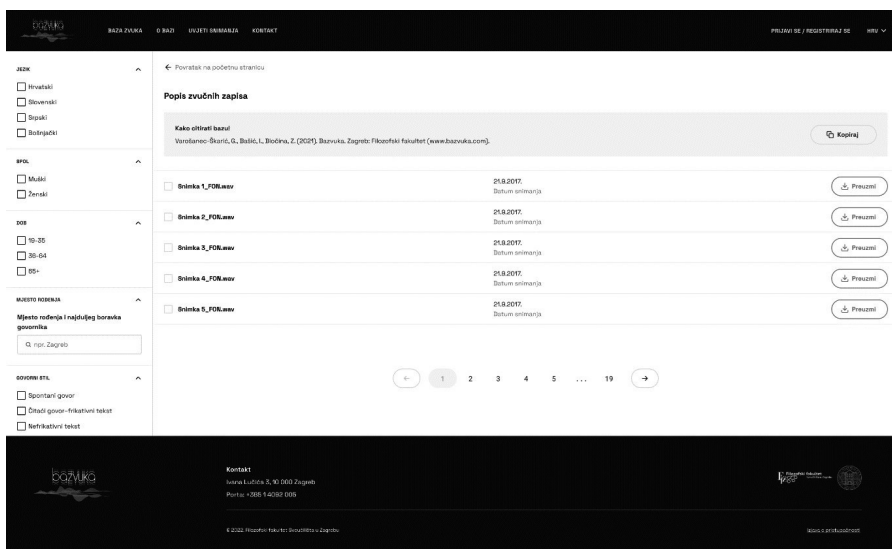
3 Zanimljiv je podatak da je za potrebe izrade dugotrajnoga prosječnog spektra govora tekstove iz djela Predraga Matvejevića *Mediterranski brevijar* (1. izdanje, 1987., Zagreb: Grafički zavod Hrvatske), prilagodio prof. Ivo Škarić 1993. g. Prilagođeni tekstovi objavljeni su na hrvatskom jeziku i engleskom prijevodu u Varošaneć-Škarić i sur. (2021: 54), a korišteni su u istraživanjima longitudinalnih usporedbi glasa, u svrhu istraživanja timbra vokalnih govornih profesionalaca (npr. Varošaneć-Škarić, 1994).

njima, s ciljem da se dobije dovoljna količina povezanoga govora bez većih uzbuđenja, kako bi snimljeni govorni materijal bio pogodan za mjerenje različitih zvučnih parametara kao što su prosječne vrijednosti i medijan F0 na temelju govora, rasponi i raspršenja F0, formantske vrijednosti naglašenih vokala itd. Primjeri pitanja navedeni su u Varošanec-Škarić (2019: 91–92). Razgovori su trajali od najkraćih oko 2 min do duljih oko 10 min, u prosječnom rasponu skupina od oko 6 do oko 9 min, kao što je navedeno u korpusu.

2.3. Dimenzije pretraživanja u zvučnoj bazi

S obzirom na podatke navedene u korpusu, metapodatci u zvučnoj bazi *Bazvuka* odnose se na korpus govornika – na društvene (jezik, razina obrazovanja, zanimanje, podrijetlo – grad, regija) i biološke informacije (dob, spol), tehničke uvjete snimanja te govorni stil. Dakle, metapodatci će se moći pretraživati s obzirom na: 1. jezik (hrvatski, srpski, bosanski, slovenski), 2. spol (muški i ženski⁴), 3. dob (mladi, srednje dobi, stariji govornici), 4. govorni stil (čitanje, spontani govor, fonacija), 5. uvjete snimanja (u studiju za zvučna snimanja, u sobi sa sniženom razinom buke), 6. razinu obrazovanja (niska, srednja, viša, visoka) te 7. zanimanje govornika. Primjer pretraživanja audiosnimki u govornoj bazi vidljiv je na slici 1.

4 Podjela s obzirom na rod, u koju bi se unijeli sociodemografski podatci te montirane audiosnimke za transrodne osobe, planirana je u budućoj zvučnoj bazi atipičnih govornika hrvatskoga jezika, za koju se prikupljaju glasovi od 2016. g.



Slika 1. Pretraživanje govornih snimki u bazi govora *Bazvuka*

Za svakoga govornika pridruženi su i podaci o mjestu rođenja, državi rođenja, državi najduljega prebivanja, mjestu najduljega prebivanja, podrijetlu oca i majke (mjesto rođenja, država rođenja), u slučaju da je to važno nekom istraživaču. Nakon rezultata pretrage, uz svaku je snimku pridružena opcija preuzimanja datoteke. Govorne je snimke moguće pretražiti bez registracije, ali za preuzimanje snimke/i nužna je registracija. Pri registraciji, korisnik se baze *Bazvuka* obvezuje postupati s preuzetim govornim snimkama u skladu s Etičkim kodeksom i pravilima korištenja baze (obvezno prihvaćanje Uvjeta korištenja, Etičkoga kodeksa i Pravila citiranja).

3. Primjer istraživanja na temelju korpusa zvučne baze „Bazvuka”

3.1. Usporedba ritmičkih mjera hrvatskoga, bosanskoga i srpskoga jezika

Govorna baza može biti polazište različitim unutarjezičnim i međujezičnim istraživanjima segmentnih i suprasegmentnih slojeva govora, ali i jezičnim te govornim analizama na višim razinama. U primjeru istraživanja u nastavku rada, autorice su željele ispitati razlikuju li se srodni jezici kao

što su hrvatski, bosanski i srpski među govornicima koji su se izjasnili da govore tim jezicima, a rođeni su i žive u Bosni i Hercegovini. Poznato je da su hrvatski, bosanski i srpski jezik južnoslavenski jezici, a brojna fonološka istraživanja pokazuju da se razlikuju na svim jezičnim razinama, tj. na leksičkoj, semantičkoj, sintaktičkoj, morfološkoj i fonološkoj (segmentna i suprasegmentna razina).

Fonološko razlikovanje zamjetno je i u slušnoj percepciji. Primjerice, razlikovanje između govornika hrvatskoga jezika i govornika srpskoga jezika najprimjetnije je u vokalskom sustavu, a govornici bosanskoga razlikuju se u ritmicitetu zbog snažne zanaglasne redukcije, trajanja naglašanih dugih slogova i zanaglasnih duljina. Razlikovanje u vokalskom sustavu između hrvatskoga i srpskoga jezika istraženo je na verificiranom manjem uzorku izvornih muških govornika (N = 28; Varošaneć-Škarić i Bašić, 2015), što je potvrđeno i na većem uzorku izvornih muških i ženskih govornika obaju jezika (N = 182; Bašić, 2018). Ranije pilot-istraživanje na manjem uzorku Varošaneć-Škarić i sur. (2017) pokazalo je da se razlikuju vokalski prostori verificiranih izvornih muških govornika hrvatskoga, srpskoga i slovenskoga jezika. Slušna procjena lingvističke bliskosti odnosno udaljenosti hrvatskoga i srpskoga jezika u procjeni govornika hrvatskoga jezika pokazala je značajne razlike u odnosu na jezičnu bliskost ($p < 0,00001$), zatim u izgovoru vokala, konsonanata, naglasaka, a najmanja je razlika, iako statistički značajna, utvrđena u razumljivosti (Varošaneć-Škarić, 2017). Nedvojbeno, međujezična razlika u kvaliteti i trajanju vokala, pogotovo u dugim pozicijama, te vrijeme trajanja vokala u usporedbi s konsonantima mogu utjecati i na razlike u mjerama ritmiciteta. Jezična osobina ritmiciteta zanimljiva je u komparativnim istraživanjima različitih jezika. Tako su Ramus i sur. (1999) analizirali razlike u jezičnom ritmu između osam različitih jezika (engleskoga, poljskoga, francuskoga, nizozemskoga, japanskoga itd.). Svrha je našega primjera pomoću različitih mjera ritmiciteta ispitati razlikuju li se jezično bliski jezici kao što su hrvatski, bosanski i srpski i u slučaju da su govornici koji se izjašnjavaju da govore tim jezicima rođeni i žive u Bosni i Hercegovini.

3.2. Postupak

U svrhu navedenoga ispitivanja iz baze je uzet u obzir korpus od 34 muška i ženska govornika koji su u demografskom upitniku izjavili da govore hrvatskim, bosanskim ili srpskim jezikom. Iz korpusa je izdvojeno devet govornika, po tri za svaki analizirani jezik. Govornici iz spomenutoga korpusa snimljeni su 2018. godine u prostorijama bez zamjetne buke na Filozofskome fakultetu u Sarajevu⁵. Metodologija snimanja bila je jednaka kod svih govornika, kvalitetnim snimačima zvuka i mikrofonima, a tijekom snimanja nadzirali su fonetičari intervjuisti. Govornicima je dana uputa da neutralnom interpretacijom bez zamjetne ekspresije pročitaju kraći tekst *Duga* (Fairbanks, 1960) u trajanju do 40 s, a potom je nastavljeno snimanje spontanoga govora u formi intervjuja s fonetičarima. Intervju sa svakim govornikom trajao je oko 10 min. Nakon montiranja govorne snimke, koje uključuje uklanjanje govora intervjuista fonetičara te eventualne šumove, u prosjeku preostaje pet minuta govora po govorniku.

Izvorni snimljeni audiozapisi montirani su u programima za akustičku analizu i obradu zvuka Praat (Boersma i Weenik, 2010), Cool Edit Pro i/ili Audacity. Uređeni govorni zapisi pohranjeni su na osobnim računalima voditeljice baze i suradnica te na repozitoriju baze *Bazvuka*. Osim tzv. čišćenja govornih snimki od nepotrebnih šumova, govorne su snimke montirane i spremljene u *wave* formi (.wav) i ovisno o govornome stilu govornika. Tako je, primjerice, govornik imao po dva govorna zapisa (npr. X7hjP⁶_CIT_DU prvi zapis za čitanje teksta *Duga* i drugi zapis za spontani govor X7hjP_SG).

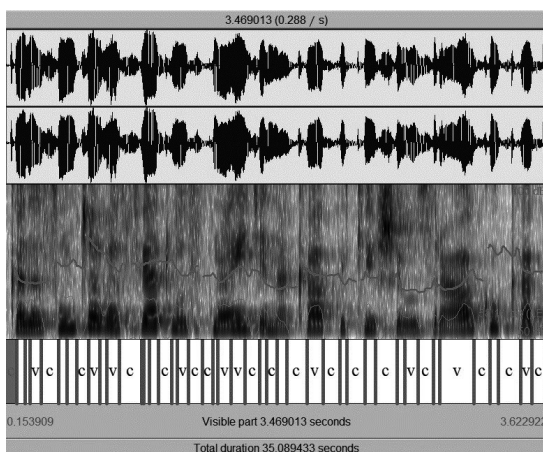
Osim audiosnimke, za potrebe baze *Bazvuka* i istraživanja mjera ritmiciteta među analiziranim jezicima u Excel dokumentu unesene su i prikupljene sociodemografske informacije, tj. metapodatci o govorniku i snimanju.

Za potrebe analize mjera ritmiciteta u Correlatorre programu (Mairano i Romano, 2010) bila je nužna prethodna anotacija govornih zapisa (slika 2), koja je provedena u Praat programu (anotacija započinje označa-

5 Govornici su snimljeni u dvjema manjim prostorijama sa sniženom razinom buke, prema postupku navedenom u metodologiji snimanja za izradu zvučne govorne baze *Bazvuka*.

6 Identitet je svakoga govornika anoniman i zaštićen jedinstvenom šifrom prema načelima Etičkoga kodeksa – npr. X7hjP za jednoga muškoga govornika hrvatskoga jezika.

vanjem govornoga zapisa koji se želi anotirati, a potom se odabiru opcije *Annotate* – to *Text Grid*, nakon kojih je nužno označavanje zvuka i formiranja *Text Grid*-a i pokretanje opcije *View and Edit*). Pri anotaciji govornih zapisa uvažene su preporuke iz rada autora Correlatorre programa (Mairano, 2011) prema kojima se dugi vokal označavao kao slijed dvaju vokala (/V/V/), trajanje početnoga dijela okluziva ograničeno je na maksimalno 70 ms (ako spektrogramski prikaz nije bio dovoljno jasan i snažan kriterij za određivanje početka okluzije), trajanje vokala određeno je prema konturama drugoga formanta (točka nestajanja F2 označavala je kraj vokala), granice poluvokala određene su prema konturama formanata, odnosno prema tranzijentima, slogotvorno [r̩] anotirano je kao vokal (/V/), obezvučeni vokal uslijed glotalizacije anotiran je kao /V/, a pauze u govoru označavale su se prazninom koja je bila omeđena kosim zagradama s lijeva i desna (/ /). Correlatorre program koristi se za analizu ritma među varijeteta istoga jezika ili, češće, za međujezična ispitivanja ritma, a temelji se na akustičkim korelatima govornoga ritma.



Slika 2. Primjer anotiranja kraćega dijela govornoga zapisa u trajanju od 3,5 s (od ukupno 35 s) jedne govornice hrvatskoga jezika u Praat programu

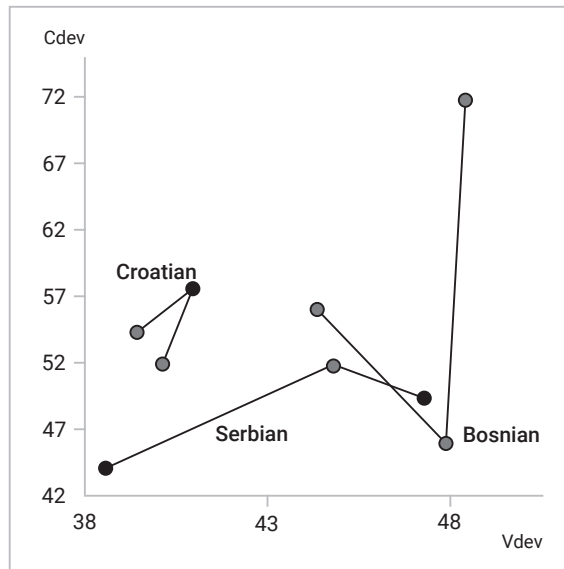
S obzirom na to da su mjere ΔC i ΔV uvelike pod utjecajem tempa govora (Ramus, 2002), za potrebe istraživanja korištene su četiri stabilnije mjere ritmiciteta: standardna devijacija suglasničkih (Cdev) i vokalskih intervala (Vdev), normalizirani indeks varijabilnosti uparenih vokalskih intervala (VnPVI) i osnovni indeks varijabilnosti uparenih vokalskih intervala

(VrPVI). PVI mjere ukazuju na varijabilnost vokalskih i konsonantskih intervala, uzimajući pritom u obzir trajanje, ali i broj intervala, u čemu Marirano (2011) vidi prednost pred drugim mjerama ritmiciteta (npr. ΔC ili ΔV). Autor također napominje da je rPVI u većoj mjeri diskriminatoran u jezicima sa složenom slogovnom strukturom u kojima se ostvaruju kompleksni konsonantski nizovi u klasterima (npr. u poljskome jeziku moguć je niz od pet konsonanata u klasteru). Spomenuta mjera računa razliku u trajanju svih parova intervala (vokala i konsonanata) u nizu te naposljetku prosjek svih razlika. S druge pak strane, normaliziranoj mjeri nPVI⁷ dodana je normalizacija koja se ogleda u dijeljenju trajanja svakoga intervala s prosječnim trajanjem parova intervala (npr. trajanje intervala xy / prosječno trajanje svih parova intervala).

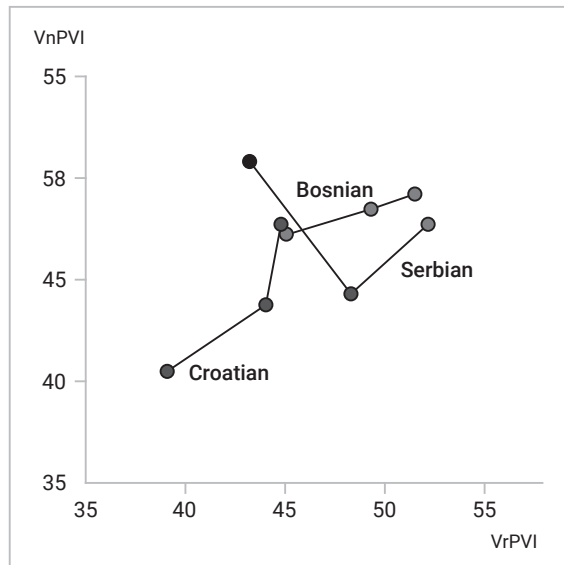
3.3. Rezultati

Rezultati ovoga pilot-istraživanja pokazali su da govornici hrvatskoga jezika ostvaruju jednostavnu slogotvornu strukturu i nižu razinu vokalske redukcije, što se očituje u vrijednostima parametara Vdev, Cdev, VnPVI i CrPVI, u usporedbi s govornicima srpskoga jezika. Zanimljivo je da govornici srpskoga jezika ostvaruju manje složenu slogotvornu strukturu od govornika hrvatskoga i bosanskoga i veću razinu vokalske redukcije od govornika hrvatskoga, ali ipak manju razinu vokalske redukcije od govornika bosanskoga. U slušnoj percepciji govornici bosanskoga jezika doživljavaju se kao govornici s najvišim stupnjem redukcije vokala u odnosu na druga dva jezika, što potvrđuju i analizirani parametri ritmiciteta. Na temelju vrijednosti mjera ritmiciteta može se zaključiti da podatci potvrđuju pretpostavku o visokoj razini slogotvorne strukture i prisustvo vokalske redukcije kod govornika bosanskoga jezika. Na slici 3 mogu se vidjeti rezultati za Cdev i Vdev, a slika 4 pokazuje rezultate za VnPVI i VrPVI. Stoga možemo zaključiti da su navedene ritmičke mjere pogodne za određivanje međugovorničkih razlika govornika srodnih jezika čak i ako su rođeni u istoj zemlji. To odražava činjenicu da su jezične granice kontinuirane i da nisu jasne kao političke nacionalne granice.

7 Normalizirana nPVI mjera daje nam podatak o razini vokalske redukcije u jeziku: što su vrijednosti nPVI mjere više, to su u jeziku vokalske redukcije veće.



Slika 3. Prikaz vrijednosti mjera Cdev i Vdev kod govornika hrvatskoga, bosanskoga i srpskoga jezika



Slika 4. Prikaz vrijednosti mjera VnPVI i VrPVI kod govornika hrvatskoga, bosanskoga i srpskoga jezika

Na slikama 5, 6 i 7 prikazane su vrijednosti različitih mjera ritmiciteta (metrike) kod hrvatskih, bosanskih i srpskih govornika.

Metrics	Values A	Values B	ErrBar A	ErrBar B
FILE	ProsjeciHrvatskenorme			
intV	476			
intC	471			
pause	56			
Vmean	88.11593622966399			
Cmean	104.30480571191002			
Vperc	46.060711328671516	0	0.82	0
Vdev	40.18916319422329	38.97776574761402	0.41	0.19
Cdev	54.622271544165436	53.65865193035614	1.60	1.57
varcoV	45.6320918320039	43.39543580145611	0.41	0.51
varcoC	52.364480731952085	53.00856752268617	1.09	2.15
Vrpvi	42.33404955079414	43.22559443056237	1.53	1.59
Crpvi	59.95200192365178	62.83002352744389	0.86	0.94
Vnpvi	44.440272391201994	44.24870659449896	1.83	1.95
Cnpvi	57.15433350905814	61.50736869608875	0.49	2.24
Vcci	38.27074843998705	39.92821097997885	1.85	1.97
Ccci	36.130751566984706	35.82739692214537	0.84	0.54
colour	#119955			
border	black			
symbol	c			

Slika 5. Prikaz vrijednosti različitih mjera ritmiciteta (metrike) kod govornika hrvatskoga jezika

Metrics	Values A	Values B	ErrBar A	ErrBar B
FILE	ProsjeciBosanskenorme			
intV	469			
intC	467			
pause	42			
Vmean	91.69398293112306			
Cmean	101.65176695731947			
Vperc	47.56950410119631	0	1.44	0
Vdev	46.90233597575078	47.074803783337835	1.27	1.08
Cdev	58.0414083233652	57.56068181016929	7.47	5.68
varcoV	51.14452385211077	49.37621363843158	0.26	0.52
varcoC	56.69843662148435	54.46363065410002	4.94	2.97
Vrpvi	47.509653932158606	52.972916841951	1.66	2.19
Crpvi	58.1256702909393	65.56830183078783	3.82	1.30
Vnpvi	48.322947593120205	51.8653433751453	0.51	1.64
Cnpvi	55.26841140885847	58.871059648855415	1.92	2.86
Vcci	36.947895544823595	39.77885931307738	5.72	7.86
Ccci	33.59041574204496	31.375681335409055	1.44	0.42
colour	#5588AA			
border	black			
symbol	c			

Slika 6. Prikaz vrijednosti različitih mjera ritmiciteta (metrike) kod govornika bosanskoga jezika

Metrics	Values A	Values B	ErrBar A	ErrBar B
FILE	ProsjeciSrpskenorme			
intV	477			
intC	466			
pause	55			
Vmean	93.00550353804302			
Cmean	95.77270423812685			
Vperc	49.716900472779734	0	1.44	0
Vdev	43.43851100757831	39.35868665968582	262	1.37
Cdev	48.16561769359174	47.092604164388234	2.40	1.61
varcoV	46.873172608258365	41.71012343129539	1.17	2.05
varcoC	50.2469467977642	50.62947321466901	0.75	0.75
Vrpvi	46.76745640388825	46.61615354274036	2.34	0.95
Crpvi	51.94347056808886	51.7641191481355	2.84	2.77
Vnpvi	47.72294213801381	47.51801680678626	1.66	2.48
Cnpvi	53.322515897118244	54.73515341346157	0.90	2.09
Vcci	42.466663263389876	42.848365830773325	2.99	1.99
Ccci	32.851034551899325	32.56011213195514	1.42	3.66
colour	#0033FF			
border	black			
symbol	c			

Slika 7. Prikaz vrijednosti različitih mjera ritmiciteta (metrike) kod govornika srpskoga jezika

Ovim smo primjerom željeli pokazati da se na temelju baze mogu provoditi međujezična istraživanja ritmiciteta srodnih jezika, a baza može biti pogodna i za istraživanja nesrodnih jezika. Bilo bi zanimljivo istražiti kojim su jezicima hrvatski, bosanski i srpski slični po ritmicitetu, tj. može li primjerice bosanski jezik zbog redukcije vokala biti u skupini jezika poput engleskoga. Nadalje, može se istraživati i unutarjezični ritmicitet npr. hrvatskoga jezika kojim bismo utvrdili razliku li se alokalni (iz)govori od dijalektalnih itd.

4. Zaključak

Prije svega, kreiranjem nove zvučne baze govora *Bazvuka* željeli smo postići dostupnost i veći opseg metapodataka znanstvenicima kojima je govor u fokusu istraživanja, posebice onima koji istražuju govor govornika hrvatskoga jezika i srodnih jezika, potom studentima, diplomantima studija fonetike, doktorandima lingvistike itd. Budući da je planirano uključivanje baze u (HR)-CLARIN infrastrukturu, baza će biti dostupna i široj znanstvenoj i stručnoj zajednici. Baza je doprinijela razvoju forenzične fonetike i sociofonetike, a u tom području mogu se na temelju većih uzoraka govornika,

posebno za hrvatski jezik, nastaviti i daljnja istraživanja. Osim što se zvučna baza govora pokazala korisnom za prepoznavanje govornika, može biti korištena i za kontrastivne analize, što smo pokušali oprimirati prigodnim ispitivanjem usporedbe mjera ritmiciteta između govornika hrvatskoga, bosanskoga i srpskoga jezika. Budući da je baza *Bazvuka* kreirana tako da bude moguća naknadna nadogradnja novim govornim snimkama, i autorima baze bit će neizostavan korpus za buduća vlastita istraživanja. Kako bi baza bila što korisnija i lingvistima, planira se i nadogradnja baze transkriptima postojećih snimki.

5. Literatura

- Audacity. (2022). <https://www.audacityteam.org/>. [posljednji pristup 20. travnja 2022.].
- Bašić, I. (2018). *Akustička analiza općeprihvaćenoga hrvatskoga i srpskog govora – formantska analiza i mjere fundamentalne frekvencije* (neobjavljen doktorski rad). Filozofski fakultet Sveučilišta u Zagrebu.
- Bašić, I. i Biočina, Z. (2020). Metodologija akustičkih istraživanja. *Zbornik radova Filozofskog fakulteta u Splitu*, 13, 19–40. <https://doi.org/10.38003/zrffs.13.10>
- Bašić, I. i Varošaneć-Škarić, G. (2019). Formant (F1-F3) and F0 gender differences between speakers of Croatian and Serbian RP. *28th Annual Conference of the International Association for Forensic Phonetics and Acoustics* (42–43). Istanbul, Turkey.
- Biočina, Z. (2019). *Prepoznavanje i akustička analiza govora otoka Brača* (neobjavljen doktorski rad). Filozofski fakultet Sveučilišta u Zagrebu.
- Biočina, Z. i Varošaneć-Škarić, G. (2017). Speaker recognition from the island of Brač. U G. Varošaneć-Škarić i A. Runjić-Stoilova (ur.), *26th Annual Conference of the International Association for Forensic Phonetics and Acoustics* (109–112). Hrvatsko filološko društvo.
- Biočina, Z., Varošaneć-Škarić, G. i Bašić, I. (2018). Prozodijski sustav Praznica. *Fluminensia*, 30(1), 103–126. <https://doi.org/10.31820/f.30.1.6>
- Boersma, P. i Weenink, D. (2010). Praat: A system for doing phonetics by computer (verzija 5.1.45). Dostupno na <http://www.fon.hum.uva.nl/praat/> [posljednji pristup 10. rujna 2022.].
- Cool Edit Pro. (2021). <https://cool-edit-pro.soft32.com/>. [posljednji pristup

14. prosinca 2021.]].
- Fairbanks, G. (1960). *Voice and articulation drillbook*, 2. izdanje. Harper & Row.
- Kišiček, G. (2012). *Forenzično profiliranje i prepoznavanje govornika prema gradskim varijetetima hrvatskoga jezika* (neobjavljen doktorski rad). Filozofski fakultet Sveučilišta u Zagrebu.
- Kišiček, G. i Varošaneć-Škarić, G. (2011). Comparison of auditory assessment of voice pitch and acoustic values of F0, formant frequencies (F1, F2, F3 and F4) and region assessment of Croatian speakers. U S. Moosmuler (ur.), *Abstracts for the 20th Annual Conference of the International Association for Forensic Phonetics and Acoustics* (34). Acoustic Research Institute.
- Kišiček, G. i Varošaneć-Škarić, G. (2016). Utjecaj kvalitete glasa na procjenu regionalnosti, socioekonomskih i psiholoških karakteristika govornika. U S. Botica, D. Nikolić, J. Tomašić i I. Vidović Bolt (ur.), *Šesti hrvatski slavistički kongres: zbornik radova*, u Vukovaru i Vinkovcima od 10. do 13. rujna 2014.: prvi svezak (309–321). Hrvatsko filološko društvo, Hrvatski slavistički odbor.
- Mairano, P. (2011). *Rhythm typology: Acoustic and perceptive studies*. *Linguistics* (doktorski rad). Università degli studi di Torino.
- Mairano, P. i Romano, A. (2010). Un confronto tra diverse metriche ritmiche usando Correlatore. U S. Schmid, M. Schwarzenbach i D. Studer (ur.), *La dimensione temporale del parlato, (Proc. of the V National AISV Congress, University of Zurich, Collegiengengebäude, 4–6 February 2009)*, Torriana (RN): EDK, 79–100. [posljednji pristup 26. travnja 2019.].
- Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. *Proceedings of Speech Prosody 2002* (115–120), Aix-en-Provence, 11–13/04/02.
- Ramus, F., Nespore, M. i Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265–292.
- Varošaneć-Škarić, G. (1994). Glasovi radio-televizijskih spikera na temelju njihova dugotrajnoga prosječnoga spektra. *Govor*, 11(1), 27–40.
- Varošaneć-Škarić, G. (2017). Assessment of linguistic closeness and comparison of acoustic characteristics of Croatian and Serbian languages. U M. Sovilj, S. Jovičić, M. Subotić i S. Maksimović (ur.), *Speech and Language 2017 Proceedings* (205–210). Life activities advancement center,

- The Institute for Experimental Phonetics and Speech Pathology.
- Varošanec-Škarić, G. (2019). *Forenzična fonetika*. Ibis grafika.
- Varošanec-Škarić, G. i Bašić, I. (2015). Acoustic characteristics of Croatian cardinal vowel formants (F1, F2 and F3). U M. Sovilj i M. Subotić (ur.), *5th International Conference on Fundamental and Applied Aspects of Speech and Language* (41–49). Life Activities Advancement Center and The Institute for Experimental Phonetics and Speech Pathology.
- Varošanec-Škarić, G., Bašić, I. i Kišiček, G. (2017). Comparison of vowel space of male speakers of Croatian, Serbian and Slovenian language. U G. Varošanec-Škarić i A. Runjić-Stoilova (ur.), *IAFPA 26 Split* (142–146). Hrvatsko filološko društvo.
- Varošanec-Škarić, G. i Kišiček, G. (2009). Izvanjske indeksikalne osobine govornika varaždinskoga i osječčkoga govora. *Suvremena lingvistika*, 35(67), 109–124.
- Varošanec-Škarić, G., Pavić, I. i Kišiček, G. (2014). Indeksi sličnosti i različitosti kod govornika hrvatskoga jezika u nefiltriranim i filtriranim uvjetima. *Suvremena lingvistika*, 40(77), 61–76.
- Varošanec-Škarić, G., Stevanović, S. i Bašić, I. (2021). Comparative perceptual evaluation and acoustic voice analysis of a transgender client male to female before and after laser-assisted voice adjustment surgery. *Hrvatska revija za rehabilitacijska istraživanja*, 57(1), 40–55. <https://doi.org/10.31299/hrri.57.1.2>

„Bazvuka” – speech database of native Croatian speakers and speakers of related languages

From 2011 to the end of 2020, 621 native speakers of Croatian, Serbian, Slovenian and Bosnian were recorded. Different speaking styles were included: reading style (reading words, sentences, short texts), spontaneous speech in the form of interviews and/or vowel phonations. Recording conditions were similar (audio recording with high-quality equipment in rooms with low noise level, 10 cm distance between mouth and microphone at an angle of 45°). A total of 45.25% of men and 54.75% of women were recorded. Approximately 84.18 hours of audio recordings were made (6.03 minutes per speaker). The Croatian language corpora was recorded in 13 Croatian cities and in 10 different locations on the island of Brač. Speakers of related languages were recorded in the countries' capitals (Belgrade, Ljubljana, and Sarajevo). The speech recordings were processed using various programmes for acoustic analysis of speech sound. The recordings were also modified for access to the infrastructure (HR-) CLARIN, which is part of the European Research Infrastructure. The database was designed to be integrated with other speech recordings in the long term. Phoneticians, linguists, students, researchers and other AAI@EduHr clients will be granted free access. This paper presents research on interlingual rhythmicity in Croatian, Bosnian, and Serbian, with audio recordings of speakers of the aforementioned languages downloaded from the *Bazvuka* database.

Keywords: *Bazvuka*, speech database, Croatian language, related language