

NAČRT RAVNANJA Z RAZISKOVALNIMI PODATKI

OBRAZEC ARIS 2.0

Pregled vsebine:

0. Splošne informacije
1. Povzetek in opis raziskovalnih podatkov
2. Shranjevanje in varnostno kopiranje podatkov
3. Zagotovitev podatkov na način FAIR prek repozitorijev
4. Etični in pravni vidiki
5. Drugi raziskovalni rezultati
6. Finančna sredstva

OBRAZEC

0	Splošne informacije	
0.1	Šifra projekta	J6-20777
0.2	Naziv projekta	EPIC-SI – Zgodnja komunikacija med starši in otroki: korpusne raziskave
0.3	Datum začetka in konca projekta	01. 03. 2026 – 28. 2. 2029
0.4	Opis projekta z vidika ravnanja z raziskovalnimi podatki	Projekt EPIC-SI vzpostavlja prvi celovit korpus zgodnje komunikacije med starši in otroki v slovenskem jeziku, saj so obstoječi viri za slovenščino omejeni in nereprezentativni. Namen projekta je vzpostaviti reprezentativen in uravnotežen korpus komunikacije med starši in otroki (12–48 mesecev) ter izvesti pilotne raziskave tipičnega jezikovnega in komunikacijskega razvoja. Glavni problem, ki ga raziskava obravnava, je pomanjkanje ustreznih, dovolj obsežnih in reprezentativnih korpusnih virov za spremljanje zgodnjega jezikovnega razvoja v slovenščini, kar omejuje razvojne, jezikoslovne in tehnološke raziskave (npr. prilagajanje orodij NLP, kot so ASR in poravnava fonemov, slovenskemu otroškemu govoru). V okviru projekta bodo zbrani novi avdio-video posnetki interakcije med otroki in starši, ki bodo prepisani in anotirani, dopolnjeni z demografskimi podatki ter obdelani v skladu z načeli FAIR. Podatki imajo osrednjo vlogo v projektu: zbiranje in obdelava avdio-video posnetkov predstavljata temeljno raziskovalno aktivnost, prepisi in anotacije pa omogočajo nadaljnjo jezikoslovno in razvojno analizo ter interpretacijo rezultatov. Podatki so neposredno povezani z dvema osrednjima ciljema projekta: (1) razvoj korpusa kot trajne podlage za nadaljnje

		korpusne raziskave in (2) izvedba pilotnih študij, ki bodo služile kot izhodišče za nadaljnje raziskave in aplikacije. Projekt ni nadaljevanje ali ponovitev predhodne raziskave, temveč ustvarja povsem nov, prvi tovrstni korpus za slovenski jezik.
0.5	Šifra vodje projekta	30369
0.6	Ime in priimek vodje projekta	dr. Amanda Saksida
0.7	Ime in priimek osebe, ki je v RO zadolžena za podporo pri ravnanju z raziskovalnimi podatki	dr. Jan Jona Javoršek, CMI IJS (pooblaščen oseba za varstvo podatkov – DPO)
0.8	Ime in priimek odgovorne osebe za ravnanje z raziskovalnimi podatki pri projektu	dr. Amanda Saksida (vodja projekta, IJS)
0.9	Elektronski naslov odgovorne osebe za ravnanje z raziskovalnimi podatki pri projektu	amanda.saksida@ijs.si
0.10	Različica in datum NRRP	1.0 (5. 6. 2026)
1	Povzetek in opis raziskovalnih podatkov	
1.1	Ali boste pri projektu ponovno uporabili že obstoječe podatke (lastne ali tuje)?	☐ Da ☒ Ne Projekt ustvarja prvi celovit korpus slovenskega otroškega jezika. Zbrani bodo novi avdio-video posnetki interakcije med otroki in starši, saj so obstoječi viri za slovenščino omejeni in nereprezentativni.
1.2	Katere vrste podatkov boste ustvarili oz. ponovno uporabili in v katerih formatih bodo shranjeni?	Ustvarjeni bodo naslednji tipi osebnih in raziskovalnih podatkov (kategorije iz Protokola za varovanje in ravnanje z osebnimi podatki projekta EPIC-SI): Identifikacijski podatki: ime, priimek, e-pošta in telefon starša/skrbnika (v kodirni listi). Eksperimentalni podatki: avdio-video posnetki interakcije starš–otrok, shranjeni pod psevdonimno kodo. Prepisi in anotacije: ročni in avtomatski prepisi posnetkov pod psevdonimno kodo. Demografski podatki: starost in spol otroka, geografska regija, pod psevdonimno kodo. Podpisana soglasja: originali v fizični obliki, ločeno od eksperimentalnih podatkov. Formati shranjevanja: Delovni avdio-video posnetki med projektom: .mpg ali .mp4 in .wav Transkripcije in metapodatki: standardiziran XML/TEI. Videoposnetki za objavo: .mpg ali .mp4 (le tisti, za katere so pridobljena privoljenja). Audio posnetki za objavo: .wav.

		Poimenovanje datotek: EP01XXX_YYYYMMDD_NN.mp4/wav/xml. Kodirne liste in demografski podatki: xlsx.
1.3	Kakšna je pričakovana velikost podatkov, ki jih nameravate ustvariti oz. ponovno uporabiti?	☐ 0–10 GB ☐ 10–100 GB ☐ 100–1000 GB ☒ >1000 GB
2	Shranjevanje in varnostno kopiranje podatkov	
2.1	Kje bodo podatki med izvajanjem projekta shranjeni in varnostno kopirani?	Posnetki in izvedeni podatki se hranijo na projektnem strežniku v podatkovnem centru IJS (fizično zavarovan prostor z nadzorom dostopa). Varnostne kopije se izdelujejo dnevno samodejno na ločen strežnik IJS. Šifriranje: AES-256 v mirovanju, TLS 1.3 v prenosu. Kodirna lista (ključ za psevdonimizacijo) se hrani ločeno: fizična kopija v zaklenjeni omarici IJS, elektronska kopija na šifriranem USB-ključku, letna varnostna kopija na ločenem šifriranem mediju na drugi fizični lokaciji (podatkovni center IJS). Kodirna lista se nikoli ne hrani v oblaku, na delovnih računalnikih, e-poštnih strežnikih ali skupnih platformah.
2.2	V katerem zaupanja vrednem repozitoriju bodo objavljeni podatki v skladu z načelom »odprti, kolikor je mogoče, zaprti, kolikor je nujno«?	1) Anonimiziran oz. psevdonimiziran korpus bo po zaključku projekta deponiran v dveh zaupanja vrednih repozitorijih: CLARIN.SI (audio-video gradivo, sedež v EU) in TalkBank – CHILDES (transkripcije in anotacije, brez medijskih datotek). Med izvajanjem projekta so podatki varovani z omejenim dostopom, dnevnim varnostnim kopiranjem in šifriranjem; ob izgubi ali poškodbi je možna obnova iz varnostnih kopij. 2) Oba repozitorija omogočata trajno skrbništvo podatkov, vključno z dodelitvijo trajnega identifikatorja (DOI/handle), zato dodatni viri za zagotavljanje trajnega skrbništva niso predvideni. Za dolgoročno hrambo se ohrani psevdonimiziran korpus (brez vizualnih identifikatorjev; glasovni posnetki ostajajo potencialno identifikabilni), namenjen javni objavi. Originalni posnetki z nezamegljenimi obrabi se hranijo na varnem strežniku IJS 5 let po zaključku projekta izključno za (i) morebitni umik soglasja, ki zahteva izolacijo in izbris posnetka, in (ii) verifikacijo prepisov in anotacij ob morebitnih revizijah; po poteku roka se nepovratno izbrišejo. Identificirajoči podatki (kodirna lista, podpisana soglasja) se ne hranijo dolgoročno: uničijo se najkasneje 5 let po zaključku projekta (predvideno 2033).
3.	Zagotovitev podatkov na način FAIR prek repozitorijev	
3.1	Ali bodo vsi podatki odprto dostopni, upoštevajoč načelo „odprti, kolikor je mogoče, in zaprti, kolikor je nujno“? Če ne bodo,	☐ Da ☒ Ne Odprto dostopen bo anonimiziran korpus (brez obrazov, psevdonimiziran). Originalni posnetki z

	pojasnite posebnosti glede dostopa.	nezamegljenimi obrazi, kodirna lista, soglasja in identifikacijski podatki niso in ne bodo javno dostopni, saj gre za osebne podatke mladoletnih udeležencev in njihovih staršev/skrbnikov. Anonimiziran oz. psevdonimiziran korpus bo javno dostopen po zaključku obdelave in obogatitve podatkov, ob depozitu v CLARIN.SI in TalkBank (CHILDES), pod licenco CC BY-SA 4.0 (CLARIN.SI). Med izvajanjem projekta je dostop omejen po vlogah: odgovorni raziskovalec in administrator IJS imata poln dostop do strežnika, ostali člani projektne skupine in zunanji prepisovalci dostopajo le do psevdonimiziranih podatkov, revizorji/Etična komisija Pedagoške fakultete UL pa le do metapodatkov na zahtevo. Po objavi vsak zunanji uporabnik dostopa prek CLARIN.SI ali TalkBank (CHILDES) in nastopa kot samostojen upravljavec podatkov. Avdio- in videoposnetki se deponirajo izključno pri CLARIN.SI (sedež v EU); pri TalkBank se deponirajo le transkripcije in anotacije brez medijev, dostop do medijev pa se zagotavlja prek mehanizma CORS s strežnika CLARIN.SI, tako da posnetki fizično ostanejo v EU in je prenos osebnih podatkov v tretjo državo izključen. Morebitne posebne zahteve (npr. izjava o namenu uporabe) se uredijo skladno s pogoji repozitorijev ob depozitu. Za uporabo anotiranega korpusa je priporočljiva dodatna dokumentacija o shemi anotacij in uporabljenih orodjih (Montreal Forced Aligner, CLASSLA-Stanza, orodja za detekcijo kretenj).
3.2	Kako boste omogočili uporabljivost vaših raziskovalnih podatkov (metapodatki, ključne besede, programska oprema, licence, drugo)?	Korpus bo opremljen z metapodatki o udeležencih na psevdonimni ravni (starost in spol otroka, geografska regija) ter o posnetkih (koda, datum, zaporedna številka). Transkripcije in metapodatki bodo objavljeni v standardiziranem formatu XML/TEI v skladu s CMDI standardom; iskanje po korpusu bo omogočeno prek vmesnika NoSketchEngine pri CLARIN.SI. Predvidene ključne besede: govorni korpus; razvoj otroškega govora; fonološki razvoj; neverbalna komunikacija; obdelava naravnega jezika (NLP). Pri pripravi podatkov in metapodatkov ne bo treba uporabiti manj poznanih ali lastnih geslovnikov oz. šifrantov. Objavljeni korpus bo spremljala dokumentacija o metodologiji zbiranja, transkripcije in anotacije ter o uporabljenih orodjih, kar omogoča ponovno uporabo. Anonimizacija (defacing, psevdonimizacija) zagotavlja skladnost z varstvom osebnih podatkov pred objavo. Anonimiziran oz. psevdonimiziran korpus bo objavljen na CLARIN.SI pod licenco Creative Commons CC BY-SA 4.0; pri depozitu v TalkBank (CHILDES) velja licenca, ki jo določi repozitorij ob depozitu.

3.3	Kakšne postopke zagotavljanja kakovosti raziskovalnih podatkov boste uporabili?	Kakovost se zagotavlja z notranjim pregledovanjem in validacijo prepisov, smernicami za transkripcijo ter usposabljanjem in nadzorom zunanjih prepisovalcev. Avtomatska orodja se izvajajo izključno na psevdonimiziranih podatkih; dnevnik obdelav se hranijo 12 mesecev, nepričakovani izhodi se dokumentirajo v Registru dogodkov.
4.	Etični in pravni vidiki	
4.1	Ali boste med izvajanjem projekta obdelovali, hranili, ustvarili ali ponovno uporabili občutljive osebne podatke oziroma posebne vrste osebnih podatkov?	☒ Da ☐ Ne Da. Projekt obdeluje osebne podatke mladoletnih udeležencev in njihovih staršev/skrbnikov, vključno z avdio-video posnetki, identifikacijskimi podatki (ime, priimek, e-pošta, telefon), prepisi in demografskimi podatki (starost in spol otroka, geografska regija). Zaradi narave avdio podatkov (človeški glas) popolna anonimizacija morda ne bo vedno dosegljiva, zato se podatki obravnavajo kot psevdonimizirani in ostajajo predmet varstva po GDPR in ZVOP-2 tudi po obdelavi. Deljenje podatkov je mogoče šele po psevdonimizaciji in obdelavi (defacing). Obdelava poteka v skladu z GDPR (Uredba EU 2016/679), ZVOP-2 in informiranim soglasjem staršev/skrbnikov; pridobljena bo odobritev institucionalne etične komisije. Obdelava in javna objava glasovnih posnetkov temeljita na pisni privolitvi (člen 6(1)(a) GDPR) ter na členu 9(2)(j) GDPR in 69. členu ZVOP-2 za namen znanstvenega raziskovanja, ob tehničnih in organizacijskih ukrepih, sorazmernih z doseženo stopnjo psevdonimizacije. Pred začetkom zbiranja podatkov odgovorni raziskovalec skupaj s pooblaščen osebo za varstvo podatkov (DPO) IJS pripravi oceno učinka v zvezi z varstvom podatkov (DPIA) po 35. členu GDPR in 24. členu ZVOP-2; ob ugotovljenem visokem preostalem tveganju se izvede predhodno posvetovanje z Informacijskim pooblaščencom RS (36. člen GDPR / 24. člen ZVOP-2). Upravljevec podatkov je Institut Jožef Stefan (Jamova 39, 1000 Ljubljana). Lastništvo in pravice do uporabe ter objave podatkov se uredijo skladno s projektno pogodbo in dogovori med sodelujočimi institucijami (IJS, PEI, PEF UL, FE UL). Po objavi korpusa zunanji uporabniki nastopajo kot samostojni upravljavci podatkov; projektna skupina IJS v razmerju do njih ni skupni upravljavec niti obdelovalec.
4.2	Kako bo med raziskavo poskrbljeno za varnost podatkov in zaščito občutljivih podatkov?	Posnetki in izvedeni podatki se hranijo na projektnem strežniku v podatkovnem centru IJS (fizično zavarovan prostor z nadzorom dostopa); varnostne kopije se izdelujejo dnevno samodejno na ločen strežnik IJS, šifriranje pa poteka po standardu AES-256 v mirovanju in TLS 1.3 v prenosu.

		Kodirna lista (ključ za psevdonimizacijo) se hrani ločeno od podatkov: fizična kopija v zaklenjeni omarici IJS, elektronska kopija na šifriranem USB-ključku ter letna varnostna kopija na ločenem šifriranem mediju na drugi fizični lokaciji (podatkovni center IJS); nikoli se ne hrani v oblaku, na delovnih računalnikih, e-poštnih strežnikih ali skupnih platformah. Uporablja se sistem psevdonimizacije (kode EPXXXXX), nadzor dostopa po vlogah in izjave o zaupnosti. Glavno tveganje je morebitna identifikacija udeležencev na podlagi glasu v avdio-video posnetkih; obvladuje se z omejenim dostopom, psevdonimizacijo, defacingom videoposnetkov pred javno objavo in strogim nadzorom dostopa do kodirne liste. V primeru kršitve varnosti osebnih podatkov je predviden postopek prijave Informacijskemu pooblaščenцу RS v 72 urah skladno s 33. členom GDPR.
5.	Drugi raziskovalni rezultati	
5.1	Ali boste poleg podatkov ustvarili ali ponovno uporabili tudi druge raziskovalne rezultate?	☒ Da ☐ Ne Poleg korpusa bodo nastali metodološki in raziskovalni rezultati: smernice za transkripcijo, prilagojena orodja NLP (poravnava fonemov, oznake besednih vrst, detekcija drž/kretenj), razvojni mejniki za slovenščino ter pilotne psiholingvistične in jezikoslovne študije. Priporočljivo je, da se tudi ti rezultati delijo in omogoča njihova uporabljivost v skladu z načeli FAIR.
6.	Finančna sredstva	
6.1	Kakšni bodo stroški ravnanja s podatki in drugimi rezultati projekta po načelih FAIR in kako bodo kriti?	Stroški ravnanja s podatki (shranjevanje na strežniku IJS, varnostno kopiranje, šifriranje, transkripcija in anotacija, anonimizacija/defacing, depozit v repozitorije) so vključeni v stroške dela in opreme projekta.