

Odprti govorni model UI za slovenščino: pregled virov, metodologija in izvedbeni načrt

umetnainteligenca.ai

(Noctocode d.o.o.)

v sodelovanju z Medobčinskim društvom slepih in slabovidnih Nova Gorica

Tehnično poročilo, avgust 2026

Izvleček

Slovenščina nima sodobnega, odprto licenciranega modela za sintezo govora. Slepí in slabovidni uporabniki so v praksi vezani na sintetizator eBralec, katerega glasovna baza iz leta 2015 obsega približno 19 ur studijskih posnetkov, zadnja objavljena arhitektura (eBralec 4, 2020) pa je hibridna korpusno-parametrična, ali na tuje oblačne storitve, ki slovenščino podpirajo le v najdražjih paketih. V poročilu popišemo javno dostopne govorne in besedilne vire za slovenščino, relevantne za sintezo, skupaj z licencami; pregledamo sodobne odprte arhitekture in predlagamo dvotirno metodologijo: lahek model razreda VITS za delovanje na napravah brez povezave ter model studijske kakovosti na osnovi tokovnega ujemanja. Osrednja ugotovitev je, da glavnina dela ni v modelih, temveč v podatkih: obstoječih 51 ur studijskih posnetkov baze ARTUR zadošča za pilot, produkcijski sistem s komercialnim licenciranjem pa zahteva nova studijska snemanja s pogodbami, ki tako rabo izrecno dovoljujejo, ter prvi sistematični zajem narečnega govora, ki ga po našem pregledu javno dostopnih virov še ni. Podamo izvedbeni načrt v štirih fazah (36 mesecev, okvirno 2,5 do 4 milijone EUR) in protokol evalvacije, ki standardne metrike (MOS, WER, UTMOS) dopolnjuje s panelom slepih in slabovidnih ocenjevalcev. Vsi navedeni viri so bili preverjeni na izvornih repozitorijih avgusta 2026.

1 Uvod

Zveza društev slepih in slabovidnih Slovenije (ZDSSS) po lastnih navedbah združuje približno štiri tisoč članov; po istem viru se v Sloveniji od osem do deset tisoč ljudi srečuje s hudo okvaro vida, z resnejšimi okvarami pa od 30 do 40 tisoč [1]. Za te uporabnike je sinteza govora primarni vmesnik do pisnih besedil: pošte, dokumentov, spleta, knjig. Kakovost sinteze zato ni stvar udobja, ampak neposredno določa dostopnost pisne kulture.

Stanje slovenskih govornih tehnologij je neenakomerno. Prepoznava govora je po projektu RSDO [2] in sistemu, opisanem v [3], primerljiva s komercialnimi sistemi. Veliki jezikovni model GaMS je slovenščini dal odprt besedilni temelj [4]. Sinteza govora pa zaostaja: glasovna baza sintetizatorja eBralec je iz leta 2015 [5], zadnja objavljena arhitektura eBralec 4 (2020) je hibridna korpusno-parametrična [6], kurirani pregled slovenskih jezikovnih tehnologij pa med govornimi viri ne navaja nobenega sodobnega odprtega modela za sintezo [7]. Sodobno nevronska kakovost za slovenščino trenutno ponuja tuja storitev ElevenLabs, izključno v modelu v3, prek oblaka, s ceno okrog 0,10 USD na 1.000 znakov in omejitvijo 5.000 znakov na zahtevo [8].

Prispevki poročila so trije: (i) preverjen popis javno dostopnih govornih korpusov, leksikalnih virov in orodij za slovenščino, relevantnih za sintezo, z licencami in oceno uporabnosti (razdelek 4); (ii) utemeljena izbira arhitektur, zasnova besedilnega sprednjega dela z naglaševanjem, podatkovna strategija za knjižni jezik in narečja ter protokol evalvacije (razdelek 5); (iii) izvedbeni načrt s fazami, časovnico, ekipo, računskimi potrebami in okvirnimi stroški (razdelek 6).

2 Obstoječe stanje in sorodna dela

2.1 Sinteza govora za slovenščino

Tabela 1 povzema obstoječe sisteme. Sintetizator eBralec (Alpineon, Amebis) je nastal v projektu Knjižnica slepih in slabovidnih, sofinanciranem iz Evropskega socialnega sklada; glasovna baza obsega približno 19 ur

studijskih posnetkov z Radia Slovenija (ženski glas 6 h 3 min, moški 5 h 33 min), akustično modeliranje je parametrično s 108-dimenzionalnimi mel-kepstralnimi vektorji [5], različica eBralec 4 (2020) pa združuje korpusni in parametrični pristop [6]. Za slepe in slabovidne je eBralec brezplačen prek ZDSSS [9]; danes ponuja glasove Nadja, Renato in Maja [12]. To je še vedno osrednji slovenski sintetizator v bralnikih zaslona.

Nevronska sinteza za slovenščino obstaja le v obliki dveh skupnostnih modelov: glas sl_SI-artur za odprtokodni sistem Piper, naučen na približno 40 urah studijskega podnabiranja baze ARTUR z enim govorcem [10, 11], ter starejši model VITS, naučen na korpusu Common Voice v okviru ograda Coqui TTS (avtor NeonGeckoCom, licenca BSD-3) [13]. Oba dokazujeta izvedljivost, nista pa sistemsko vzdrževana, nimata naglaševalnika na ravni celotnega jezika in ne pokrivata več glasov.

Tabela 1: Sinteza govora za slovenščino, stanje avgust 2026. Viri: [5, 6, 8, 9, 10, 12, 13, 14].

Sistem	Leto	Tehnologija	Ključne omejitve
GOVOREC (UL FE, Amebis)	1999-2002	difonska/enotna sinteza	zastarelo; zgodovinski pomen
eBralec (Alpineon, Amebis)	2015; v4 2020	korpusno-parametrična (hibridna)	prednevronska kakovost; ~19 h posnetkov; brez narečij; za slepe brezplačen prek ZDSSS
Piper sl_SI-artur (skupnost)	2023	nevronska (VITS)	en glas; ~40 h; brez naglaševalnika; ni vzdrževan
Coqui sl/cv/vits (skupnost)	2022	nevronska (VITS)	učen na Common Voice; osnovna kakovost; ni vzdrževan
ElevenLabs v3 (tuja storitev)	2025-2026	sodobna nevrnska	samo oblak; ~0,10 USD/1.000 znakov; 5.000 znakov/zahtevo; brez jamstev za slovenščino
Microsoft Azure sl-SI	n. p.	nevronska	glasova sl-SI-PetraNeural in sl-SI-RokNeural; oblak, naročnina; brez narečij

2.2 Prepoznavanje govora za slovenščino

Prepoznavanje je razmeroma dobro pokrita; program je ne gradi znova, jo pa uporablja za poravnavo podatkov in evalvacijo. Projekt RSDO je izdal odprt konformerski model E2E [15] in govorno bazo ARTUR [16]. Sistem iz [3] na evalvaciji SloBench (15 posnetkov, skupaj približno 3,5 ure) dosega stopnjo napake na besedah okrog 5,0 %. Slovenščino podpirata tudi odprta večjezična sistema Whisper (licenca MIT) [17] in Omnilingual ASR (Apache 2.0, oznaka slv_Latn) [18]; javno je dostopen tudi Whisper-small, doučen na bazi ARTUR (WER 11,0; Apache 2.0) [19].

2.3 Sodobne odprte arhitekture za sintezo

Tabela 2 povzema kandidatne arhitekture in licence. Nobena od pregledanih odprtih večjezičnih arhitektur slovenščine izrecno ne podpira s preverjeno kakovostjo, zato jezika ne dobimo z izbiro modela; razlika med kandidati je v licenci, računski zahtevnosti in primernosti za tek na napravah. Za lahki tir je referenčna arhitektura VITS [20], ki jo uporablja Piper [10]. Za studijski tir sta referenčni arhitekturi na osnovi tokovnega ujemanja Matcha-TTS [21] in F5-TTS [22]; slednja poenostavlja pristop E2 TTS [23]. Ključno je ločiti licenco kode od licence uteži: koda F5-TTS je pod licenco MIT, objavljene uteži pa pod CC BY-NC 4.0 in so za komercialno licenciranje neuporabne; model je treba naučiti na lastnih in odprtih podatkih. Posebnost je ograda IMS-Toucan (Apache 2.0) z objavljenim večjezičnim modelom ToucanTTS, ki cilja na več kot 7.000 jezikov z metaučanjem [24]; učen je med drugim na korpusu FLEURS, ki slovenščino vključuje, neposredne potrditve slovenščine v objavljenem seznamu jezikov pa v okviru te redakcije nismo opravili, zato ga obravnavamo kot obetavno izhodišče pilota, ne kot preverjeno rešitev.

Tabela 2: Kandidatne odprte arhitekture, stanje avgust 2026. Licence preverjene na izvornih repozitorijih.

Model	Arhitektura	Licenca (koda/uteži)	Slovenščina	Vloga v projektu
VITS [20]	VAE + GAN, E2E	MIT / uteži razne	ne	osnova tira A (naprave)
Piper [10]	VITS	prej MIT, zdaj GPL-3.0	da (1 glas, skupnost)	izhodišče pilota
Matcha-TTS [21]	tokovno ujemanje	MIT	ne	osnova tira B
F5-TTS [22]	tokovno ujemanje	MIT / CC BY-NC	ne	arhitektura da, uteži ne
IMS-Toucan [24]	FastSpeech-2 razred, metaučenje	Apache 2.0 / Apache 2.0	prek FLEURS; nepotrjeno	alternativno izhodišče pilota
StyleTTS 2 [25]	difuzija + GAN	MIT / posebni pogoji	ne	primerjava
XTTS-v2 [26]	avtoresivna, kloniranje	CPML (nekomercialna)	ne (17 jezikov)	izključen (licenca)
CosyVoice 2/3 [27]	LLM + tokovno ujemanje	Apache 2.0	ne (9 jezikov)	arhitekturna referenca
Kokoro [28]	StyleTTS-2 razred, 82M	Apache 2.0	ne (9 jezikov)	dokaz majhnih modelov
MMS-TTS [29]	VITS razred, 1.107 jezikov	CC BY-NC 4.0	navedena v projektu	izključen (licenca)

2.4 Umestitev v javne politike

Resolucija o nacionalnem programu za jezikovno politiko 2021-2025 med cilje izrecno uvršča razvoj kakovostne govorne sinteze in prepoznavne, tudi za slepe, slabovidne in gluhoslepe osebe [30]. Nacionalna strategija za razvoj umetne inteligence do leta 2030 napoveduje nacionalni podatkovni prostor za slovenščino [31], Slovenija pa je polnopravna članica evropskega konzorcija za jezikovne tehnologije ALT-EDIC [32]. Projekt PoVeJMo (skupna vrednost okrog 4 milijone EUR, sredstva za okrevanje in odpornost prek agencije ARIS) je izdal odprti besedilni model GaMS; govorne sinteze ni pokril [4]. Predlagani program je torej izvedbeno nadaljevanje sprejetih usmeritev, ne nova politika.

3 Jezikoslovna izhodišča

3.1 Naglas

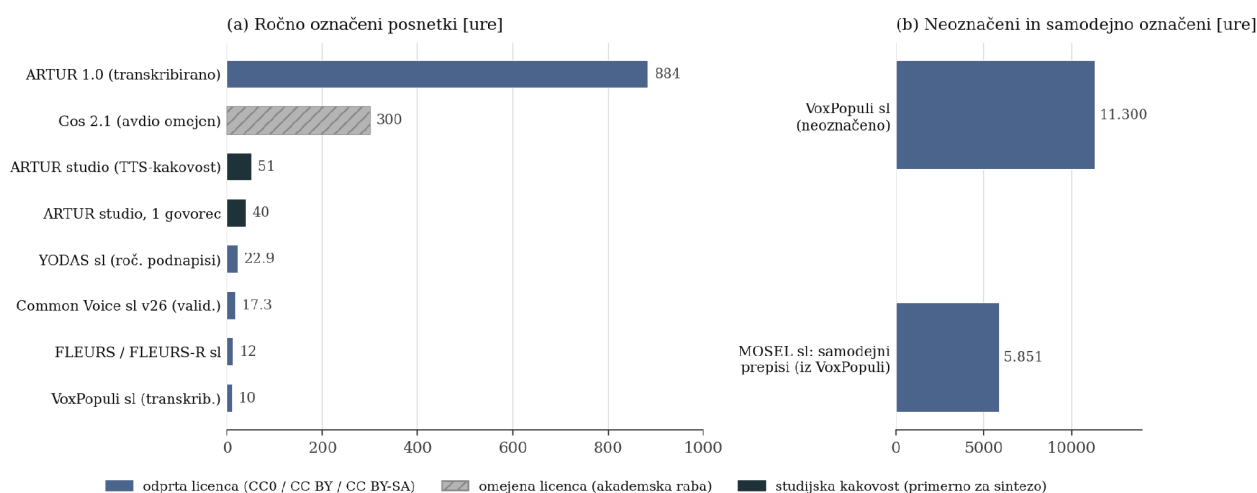
Slovenščina ima prosto naglasno mesto, ki v standardnem zapisu ni označeno; sintetizator ga mora napovedati sam, napačen naglas pa je za poslušalca najbolj moteča napaka strojnega branja. Osrednji odprti vir je leksikon Sloleks 3.1 (februar 2026): 372.341 gesel z onaglašeni oblikami (jedro približno 100.000 gesel je ročno pregledano) ter samodejno generiranimi fonetičnim prepisoma IPA in SAMPA, pod licenco CC BY-SA 4.0 [33]. Kot referenco natančnosti strojnega naglaševanja navajamo samodejno onaglašeni Sloleks 1.2, kjer je nevronska mreža dosegla 91,5-odstotno točnost naglasnega mesta oziroma 88,5-odstotno točnost mesta in tipa [34]; vir je pod nekomercialno licenco in v produkcijski verigi ne more sodelovati, potrjuje pa izvedljivost. Obsežni izgovorni leksikon SI-PRON (Alpineon) je opisan v literaturi [35], vendar javno ni dostopen, zato je odprta hrbtenica naglaševanja in grafemsko-fonemske pretvorbe Sloleks.

3.2 Narečja

Po klasifikaciji Logarja in Riglerja (1983) ima slovenščina v sedmih narečnih skupinah 37 narečij in 12 podnarečij [36]. Za sintezo to pomeni izrazito variabilnost naglasa, vokalizma in intonacije. Ključna ugotovitev popisa: narečni govorni korpus, primeren za učenje modelov, po našem pregledu ne obstaja; razpoložljivi viri so dokumentacijski, med njimi interaktivna karta narečnih besedil [37], leksikalni narečni modul portala Franček [38] in besedilna zbirka DIALECT-COPA za cerkljansko narečje [39]. Sistematični zajem narečnega govora je zato izvirni prispevek programa in hkrati njegov najdražji podatkovni del.

4 Razpoložljivi podatkovni viri

Popis je bil opravljen avgusta 2026 s preverjanjem izvornih repozitorijev (CLARIN.SI, Hugging Face, GitHub, uradne statistike Mozilla Common Voice). Slika 1 povzema govorne vire, tabela 3 podaja podrobnosti.



Slika 1: Javno dostopni govorni posnetki za slovenščino: (a) ročno označeni, (b) neoznačeni oziroma samodejno označeni. Barva označuje licenčni razred. MOSEL je izpeljan iz posnetkov VoxPopuli, zato se količini ne seštevata. Viri: [11, 16, 40, 42, 43, 44, 45, 46].

Tabela 3: Govorni korpusi z relevantno za slovenščino, stanje avgust 2026.

Vir	Obseg	Govorci	Licenca (avdio)	Primernost
ARTUR 1.0 [16]	1.067 h (884 h transkrib.); studio 51 h; parlament 201 h	~1.000	CC BY-SA 4.0	osrednji vir; studio del TTS-kakovost
ARTUR studio, podnabor [11]	40 h; 22.796 posnetkov; 22,05 kHz	1 (G0911)	izvorno CC BY-SA 4.0	takoj uporaben za pilot
Gos 2.1 [40]	~300 h; 2,46 mio besed	raznoliko	avdio omejen (akad.); transkripcije CC BY-SA	avdio za odprti model neuporaben
ROG 1.1 [41]	98.393 pojavnic; prozodična anotacija	izbor iz Gos	CC BY-SA 4.0	edini z anotirano prozodijo
Common Voice sl v26.0 [42]	21,6 h posnetih; 17,3 h valid.	1.003	CC0	evalvacija, robustnost
VoxPopuli sl [43]	4.400-11.300 h neoznač.; 10 h transkrib.	45 (transkrib.)	CC0	samonadzorno predučenje
MOSEL sl [44]	5.851 h samodejnih prepisov (iz VoxPopuli)	n. p.	CC BY 4.0	največji odprti označeni nabor; kakovost prepisov samodejna
YODAS sl [45]	22,9 h (ročni podnapisi); YODAS2 24 kHz	raznoliko	CC BY 3.0	robustnost; splet
FLEURS / FLEURS-R sl [46, 47]	~12 h; FLEURS-R zvočno obnovljen za sintezo	malo	CC BY 4.0	evalvacija; pomožni vir za sintezo

4.1 Česa ni

Preverili smo tudi vire, ki se v tovrstnih načrtih pogosto predpostavljajo, a za slovenščino ne obstajajo: zbirke ParlaSpeech za slovenščino ni (parlamentarni govor je znotraj ARTUR-P), na repozitoriju OpenSLR ni nobenega slovenskega vira [48], zbirke CMU Wilderness in MediaSpeech slovenščine ne pokrivata, slovenskih zvočnih knjig na platformi LibriVox praktično ni, za zbirko Common Voice Spontaneous Speech pa vključenost slovenščine ni potrjena. Uradnih akustičnih modelov Montreal Forced Aligner za slovenščino ni; obstaja pa odprta slovenska rešitev za prisilno poravnavo [53]. Narečnega govornega korpusa, kot rečeno, nismo našli.

4.2 Besedilni viri in gradniki sprednjega dela

- **Sloleks 3.1** [33]: 372.341 gesel, onaglašene oblike, IPA in SAMPA; CC BY-SA 4.0. Hrbtenica naglaševanja in grafemsko-fonemske pretvorbe.
- **Orodja RSDO za normalizacijo**: normalizator, denormalizator in vstavljalnik ločil, objavljeni kot odprti modeli [51]. Sprednjega dela ni treba graditi iz nič.

- **Cevovod CLASSLA** (Apache 2.0) za tokenizacijo, oblikoskladenjsko označevanje in lematizacijo standardne, nestandardne in govorne slovenščine; ista organizacija objavlja tudi zvočni razvrščevalnik primarnega naglasa na ravni okvirjev, uporaben pri pripravi podatkov [52].
- **Gigafida 2.0** [49]: 1,13 milijarde besed referenčnega korpusa za frekvenčne statistike; prosto snemljiv je le podkorpus ccGigafida (~100 mio besed) z nekomercialno licenco, kar je treba upoštevati pri gradnji učnih besedil.
- **espeak-ng** podpira slovenščino in služi kot zasilna grafemsko-fonemska pretvorba; koda je pod GPLv3, zato produkcijska veriga predvidi lasten modul iz Sloleksa [50].
- **Skupnostna orodja** (Hugging Face): eksperimentalni naglaševalnik in modela ByT5 za grafemsko-fonemsko pretvorbo slovenščine ter omenjeni glas za Piper [11]. Dokazujejo izvedljivost, dokumentacija je skopa.

4.3 Licenčna analiza

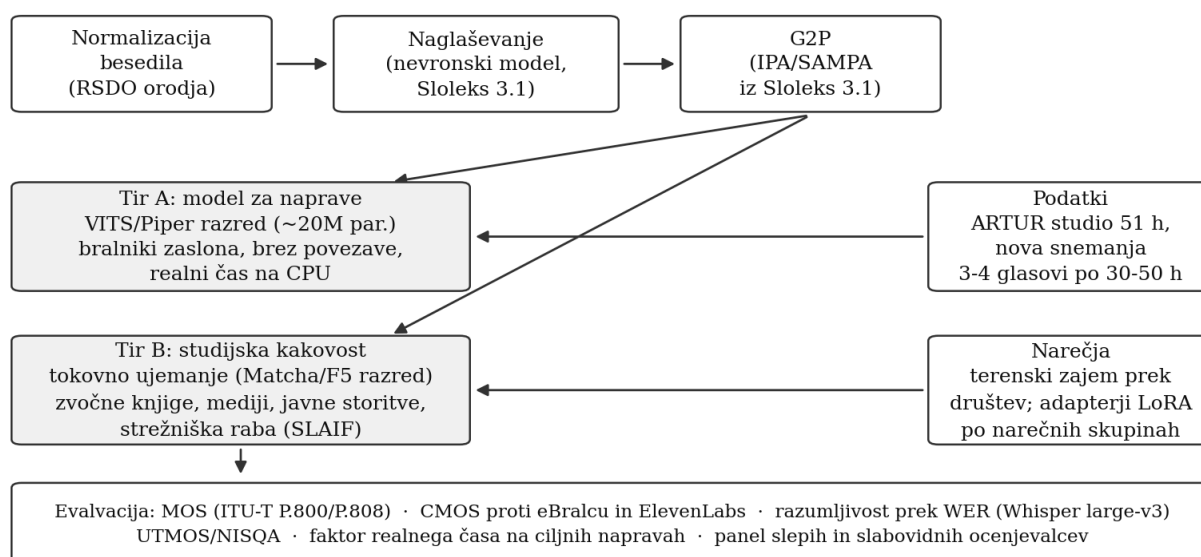
Za javni model z brezplačno rabo za prikrajšane skupine in licenčnim dostopom za razvijalce so združljivi viri pod CC0, CC BY in CC BY-SA; pri CC BY-SA je treba pravno presoditi obseg obveznosti deljenja pod enakimi pogoji v razmerju do uteži modela in izhodnega govora. Izključeni so viri z nekomercialnimi licencami (avdio Gos, MMS-TTS, objavljene uteži F5-TTS, samodejno onaglašeni Sloleks 1.2) in viri z akademsko omejenim dostopom. Nova studijska snemanja morajo imeti pogodbe z govorniki, ki izrecno dovoljujejo javno objavo modela in komercialno licenciranje; to je ceneje zagotoviti vnaprej kot popravljati za nazaj.

5 Metodologija

5.1 Dvotirna arhitektura

Zahtevi po delovanju brez povezave na skromni opremi in po studijski kakovosti za daljša besedila si nasprotujeta, zato ju ločimo v dva tira s skupnim besedilnim sprednjim delom (slika 2). Tir A je lahek model razreda VITS (red velikosti 20 milijonov parametrov) za bralnike zaslona: tek v realnem času na procesorju, brez povezave, zasebnost po zasnovi. Tir B je model na osnovi tokovnega ujemanja (razred Matcha-TTS oziroma F5-TTS), naučen na licenčno čistih podatkih, za zvočne knjige, medije in javne storitve prek strežniške rabe. Ločitev znižuje tveganje: pilot tira A je izvedljiv takoj na obstoječih podatkih, tir B raste z novimi snemanji.

Besedilni sprednji del (skupen obema tiroma)



Slika 2: Zasnova sistema: skupni besedilni sprednji del, sintezna tira A in B, podatkovni viri in evalvacija.

5.2 Besedilni sprednji del

Sprednji del obsega normalizacijo (števila, datumi, kratice, simboli), naglaševanje in grafemsko-fonemsko pretvorbo. Za normalizacijo se uporabijo in nadgradijo obstoječa orodja RSDO [51]. Naglaševalnik in pretvorba

se uči na Sloleksu 3.1 [33]; ciljna metrika je preseči objavljeno referenco 91,5-odstotne točnosti naglasnega mesta [34], merjeno na izločenem testnem delu leksikona in na namenskem naboru težkih primerov (prevzete besede, lastna imena, homografi). Sprednji del je skupen obema tiroma, kar zagotavlja enotno izgovarjavo, in se objavi kot samostojna odprta knjižnica, uporabna tudi zunaj sinteze.

5.3 Podatkovna strategija

- **Pilot na obstoječih podatkih.** Fino uglaševanje tira A na 40-urnem študijskem podnaboru ARTUR [11]; vzporedno preizkus večjezičnega modela ToucanTTS kot alternativnega izhodišča [24]. Referenčne točke izvedljivosti iz literature: prilagoditev večjezičnega modela na nov jezik je bila demonstrirana s 30 minutami parnih podatkov [54], prilagoditev govorcu znotraj podprtega jezika z manj kot minuto [55]; za produkcijsko kakovost enega glasu računamo z 20 do 50 urami študijskih posnetkov, kar je skladno z obema slovenskima referencama (eBralec približno 6 h na glas ob nižji ciljni kakovosti [5], Piper artur 40 h [11]).
- **Nova študijska snemanja.** Tri do štirje glasovi (uravnoreženo po spolu, vsaj en mlajši in en starejši glas) po 30 do 50 ur: fonetično uravnorežene povedi, daljši odstavki za prozodijo, vprašanja, števila, kratice. Snemanje v študijski kakovosti pri 48 kHz; transkripcija in prisilna poravnava z odprto slovensko rešitvijo na osnovi MFA [53] ob podpori razpoznavalnika [15]. Pogodbe z govorniki izrecno dovoljujejo javni model in komercialno licenciranje.
- **Narečni zajem.** Ciljno 5 do 10 ur uporabnega govora na narečno skupino v prvem krogu (sedem skupin [36]), prek mreže društev in univerz: vodeni pogovori in branje lokalnih besedil, dvojna transkripcija (standardizirana in poenostavljena narečna), pri pripravi si pomagamo z razvrščevalnikom naglasa [52]. Narečja modeliramo kot prilagoditvene module LoRA [56] nad skupnim modelom, kar loči strošek osnovnega modela od postopnega dodajanja narečij.
- **Neoznačeni in samodejno označeni podatki.** VoxPopuli (do 11.300 ur, CC0) [43] za samonadzorno predučenje reprezentacij; MOSEL (5.851 ur samodejnih prepisov, CC BY 4.0) [44] za robustno učenje razpoznavne in filtriran izbor dodatnih parnih podatkov; YODAS [45] in FLEURS-R [47] kot pomožna vira. Zaradi kanalske neenotnosti ti viri niso primerni za neposredno učenje študijske sinteze.

5.4 Učenje in računske potrebe

Referenčna točka: osnovni model F5-TTS je bil naučen na približno 95.000 urah govora v nekaj več kot tednu dni na osmih grafičnih procesorjih A100 80 GB (1,2 milijona korakov) [22]. Naš parni korpus bo za tri velikostne razrede manjši (150 do 250 ur), model ustrezno manjši, učni teki krajši; ob rezervi za iskanje hiperparametrov, ponovitve in narečne adapterje ocenjujemo skupno porabo programa na 10.000 do 50.000 GPU ur. To je znotraj brezplačnega dostopa, ki ga malim podjetjem in raziskovalnim organizacijam ponuja slovenska tovarna umetne inteligence SLAIF (dostop Fast Lane sega do 50.000 GPU ur) [57]. Računska infrastruktura torej ni ozko grlo in ne pomemben strošek programa.

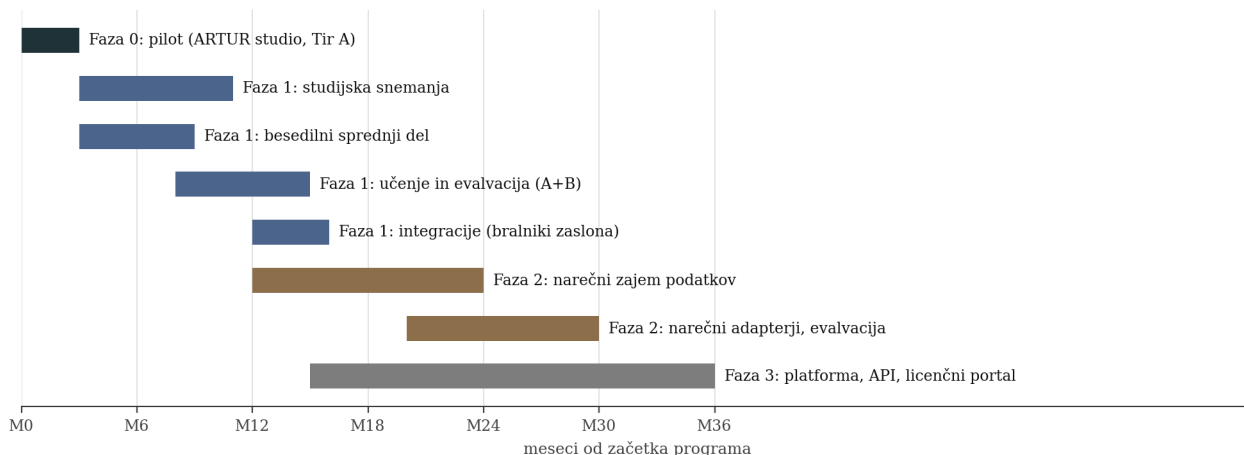
5.5 Evalvacija

- **Subjektivna kakovost.** MOS po metodi absolutnega ocenjevanja (ITU-T P.800), pri množičnem ocenjevanju po ITU-T P.808 [58]; primerjalni CMOS proti eBralcu in ElevenLabs v3 na istih besedilih.
- **Razumljivost.** WER razpoznavne sintetiziranega govora z modelom Whisper large-v3, po praksi iz literature [22]; dodatno s slovenskim razpoznavalnikom [15] zaradi občutljivosti na slovenske foneme.
- **Samodejne ocene.** UTMOS [59] in NISQA [60] za sprotno spremljanje razvoja; obe metriki sta kazalnika, ne nadomestek subjektivnih testov.
- **Panel ciljnih uporabnikov.** Slepi in slabovidni ocenjevalci prek društev: poleg MOS tudi preizkus dolgotrajnega poslušanja (30 in več minut) ter razumljivosti pri povečani hitrosti branja, ki sta za uporabnike bralnikov odločilna, standardne kampanje pa ju ne merijo.
- **Hitrost.** Faktor realnega časa tira A na ciljnih napravah (pisarniški računalnik brez grafičnega pospeševalnika, telefon srednjega razreda).

Ker javnega merila za slovensko sintezo ni, se evalvacijski nabori in protokol objavijo kot javni testni sklop s predlogom vključitve v platformo SloBench [61]; s tem postane merljiv tudi napredek prihodnjih sistemov drugih izvajalcev.

6 Izvedbeni načrt

Program obsega štiri faze v okvirno 36 mesecih (slika 3). Faza 0 je pilot omejenega obsega, ki pred polnim financiranjem programa zagotovi zvočni dokaz izvedljivosti in primerjalno oceno z obstoječimi sistemi. Priprava pričujoče raziskave in specifikacije je bila izvedena kot lastni vložek pripravljavca; izvedbene faze, vključno s pilotom, se financirajo iz programa.



Slika 3: Okvirna časovnica programa po fazah in delovnih sklopih (M označuje mesec od začetka).

Tabela 4: Faze programa z mejniki. Ocene so informativne in ne vključujejo DDV; členitev faze 1 je v tabeli 5.

Faza	Vsebina in mejnik	Trajanje	Okvirni strošek (EUR)
0	Pilot: tir A na obstoječem podnaboru ARTUR; javna demonstracija in slepa primerjava z eBralcem in ElevenLabs. Mejnik M3: zvočni dokaz.	M1-M3	80.000-150.000
1	Osnovni model: nova snemanja, sprednji del, učenje obeh tirov, integracije za bralnike zaslona. Mejnik M15: javna izdaja modela v1 z odprtimi utežmi.	M4-M15	1.100.000-1.700.000
2	Narečja: terenski zajem po skupinah, adapterji, evalvacija. Mejnika M24 in M30: prvi adapterji; pokrite vse skupine (prvi krog).	M12-M30	800.000-1.400.000
3	Platforma: javni vmesnik API, licenčni portal, vzdrževanje, nadaljnja snemanja. Mejnik M36: platforma v polnem delovanju.	M15 naprej	250.000-350.000 letno

Tabela 5: Okvirna členitev stroškov faze 1. Računska infrastruktura je predvidena prek dostopa SLAIF [57] in ni vključena.

Postavka	Ocena od (EUR)	Ocena do (EUR)
Studijska snemanja, transkripcija, poravnava (3-4 glasovi po 30-50 h)	200.000	350.000
Besedilni sprednji del (normalizacija, naglaševanje, pretvorba)	180.000	280.000
Učenje modelov in eksperimenti (tira A in B)	450.000	650.000
Evalvacija (kampanje MOS, panel uporabnikov)	60.000	120.000
Integracije za bralnike zaslona in dokumentacija	120.000	200.000
Vodenje programa, pravni pregled (pogodbe, licence)	90.000	100.000
Skupaj	1.100.000	1.700.000

6.1 Ekipa

Jedrna ekipa faze 1 obsega pet do sedem oseb: vodjo programa, dva raziskovalca oziroma inženirja za govorne tehnologije, inženirja za podatke, jezikoslovca fonetika (naglaševanje, nadzor transkripcij, v fazi 2 sodelovanje z dialektologi), inženirja za integracije ter koordinatorskega evalvacije v sodelovanju z društvi. Fazi 2 se pridružijo terenski sodelavci iz lokalnih društev in univerz.

6.2 Tveganja

Tabela 6: Glavna tveganja in blažitve.

Tveganje	Ocena	Blažitev
Pravni domet CC BY-SA glede uteži modela	srednja	pravno mnenje pred fazo 1; jedro produkcijskih glasov iz novih snemanj s pogodbami
Razpoložljivost govorcev in studia	nizka	sodelovanje s profesionalnimi napovedovalci; snemanja razporejena čez šest mesecev
Konsistentnost narečnih transkripcij	visoka	dvojna transkripcija, dialektološki nadzor, ožji prvi krog, adapterji namesto ločenih modelov
Sprejetje pri uporabnikih bralnikov	srednja	panel od pilota naprej; merjenje dolgotrajnega poslušanja in visokih hitrosti; združljivost z navadami
Vzdrževanje po koncu financiranja	srednja	licenčni prihodki platforme; odprte uteži in dokumentacija omogočajo prevzem drugemu izvajalcu

7 Omejitve in odprta vprašanja

- Sinteza slovenskih narečij je neraziskana; ocene obsega narečnih podatkov in kakovosti adapterjev so negotove in se preverijo v prvem krogu zajema.
- Ocene stroškov in računskih potreb so informativne; temeljijo na objavljenih referenčnih točkah [5, 22] in primerljivih projektih ter zahtevajo konkretizacijo obsega.
- Javnega merila za slovensko sintezo ni; primerjave s tujimi sistemi so mogoče le z lastnimi testi, kar naslavlja predlog javnega testnega sklopa (razdelek 5.5).
- Trditve tretjih projektov, ki jih nismo mogli neodvisno preveriti, so označene v besedilu: pokritost slovenščine v objavljenih glasovih MMS [29] in v modelu ToucanTTS [24] navajamo kot navedbi projektov.
- Resolucija o jezikovni politiki za obdobje po letu 2025 ob zaključku redakcije (avgust 2026) še ni bila sprejeta; sklic [30] se nanaša na resolucijo 2021-2025.

8 Zaključek

Gradniki odprtega slovenskega govornega modela obstajajo: 51 ur studijskih posnetkov in 884 ur transkribiranega govora pod odprto licenco, več tisoč ur neoznačenega in samodejno označenega govora, leksikon z onaglašeni oblikami in fonetičnimi prepisi, odprta orodja za normalizacijo in poravnavo, zrele odprte arhitekture z ustreznimi licencami kode ter brezplačna nacionalna računska infrastruktura. Manjkajo tri stvari: nova studijska snemanja s pogodbami, ki dovoljujejo javno in komercialno rabo, prvi sistematični zajem narečnega govora in institucionalni okvir, ki model vzpostavi kot javno infrastrukturo z brezplačnim dostopom za prikrašane skupine. Nobena od teh treh stvari ni tehnično tvegana; vse tri so vprašanje odločitve in organizacije. Pilot faze 0 je na obstoječih podatkih izvedljiv v treh mesecih in razpravo o financiranju opremi z zvočnim dokazom.

Zahvala. Medobčinskemu društvu slepih in slabovidnih Nova Gorica in Zvezi društev slepih in slabovidnih Slovenije za uporabniška izhodišča. Poročilo je delovno gradivo za razpravo in ne prejudicira izvajalca programa.

Literatura

- [1] Zveza društev slepih in slabovidnih Slovenije. Pogosta vprašanja.
<https://www.zveza-slepih.si/okvare-vida/pogosta-vprasanja/>, dostop 10. 8. 2026.
- [2] Projekt RSDO: Razvoj slovenščine v digitalnem okolju, govorne tehnologije.
<https://rsdo.slovenscina.eu/en/speech-technologies>, dostop 10. 8. 2026.
- [3] Govori.si: konformerski razpoznavnik slovenščine (UL FRI); rezultati na SloBench ASR.
<https://lgm.fri.uni-lj.si/wp-content/uploads/2026/03/212711427.pdf>, dostop 10. 8. 2026.
- [4] PoVeJMo: Prilagodljiva obdelava naravnega jezika s pomočjo velikih jezikovnih modelov; model GaMS.
https://reforms-investments.ec.europa.eu/projects/povejmo-adaptive-natural-language-processing-large-language-models-co-financing-research-innovation_en;
<https://www.cjvt.si/blog/slovenija-dobila-veliki-jezikovni-model-za-slovenscino-gams/>, dostop 12. 8. 2026.

- [5] J. Žganec Gros in sod. eBralec: sintetizator govora za slovenščino. Zbornik konference JTDH, 2016.
https://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH2016_zganec_gros_et_al.pdf.
- [6] eBralec 4: hibridni sintetizator govora. Zbornik konference Informacijska družba (IS), 2020.
https://is.ijs.si/wp-content/uploads/2020/10/04-HCI-IS_2020_paper_4.pdf.
- [7] CLARIN.SI. Slovenian Language Technologies Overview.
<https://github.com/clarinsi/Slovenian-Language-Technologies-Overview>, dostop 12. 8. 2026.
- [8] ElevenLabs. Models; API pricing. <https://elevenlabs.io/docs/overview/models>;
<https://elevenlabs.io/pricing/api>, dostop 12. 8. 2026.
- [9] ZDSSS. eBralec: brezplačne licence za slepe in slabovidne. <https://www.zveza-slepih.si/pripomocki/ebralec/>, dostop 12. 8. 2026.
- [10] Piper TTS. <https://github.com/rhasspy/piper> (MIT; arhiviran oktobra 2025); naslednik <https://github.com/OHF-Voice/piper1-gpl> (GPL-3.0); slovenski glas sl_SI-artur: <https://huggingface.co/rhasspy/piper-voices>.
- [11] P. Pišljari. Podnabor ARTUR studio za sintezo (en govorec, 40 h) in eksperimentalna orodja za slovenščino. https://huggingface.co/datasets/ppisljar/artur_studio_tts; <https://huggingface.co/ppisljar>, dostop 12. 8. 2026.
- [12] Amebis. eBralec: glasovi in storitve. <https://www.amebis.si/ebralec>, dostop 13. 8. 2026.
- [13] Coqui TTS, register modelov: tts_models/sl/cv/vits (avtor NeonGeckoCom, BSD-3).
<https://raw.githubusercontent.com/coqui-ai/TTS/dev/TTS/.models.json>, dostop 13. 8. 2026.
- [14] Microsoft. Azure AI Speech: language and voice support (sl-SI: PetraNeural, RokNeural).
<https://learn.microsoft.com/en-us/azure/ai-services/speech-service/language-support?tabs=tts>, dostop 13. 8. 2026.
- [15] Slovene Conformer CTC BPE E2E ASR model RSDO-DS2-ASR-E2E 2.0. CLARIN.SI.
<http://hdl.handle.net/11356/1737>; https://github.com/clarinsi/Slovene_ASR_e2e.
- [16] Govorna baza ARTUR 1.0. CLARIN.SI: posnetki <http://hdl.handle.net/11356/1776>; transkripcije <http://hdl.handle.net/11356/1772> (CC BY-SA 4.0).
- [17] A. Radford in sod. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356, 2022. Koda in uteži MIT.
- [18] Omnilingual ASR team (Meta). Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages. arXiv:2511.09690, 2025. Apache 2.0. <https://github.com/facebookresearch/omnilingual-asr>.
- [19] samolego/whisper-small-slovenian: Whisper-small, doučen na ARTUR 1.0 (WER 11,0; Apache 2.0).
<https://huggingface.co/samolego/whisper-small-slovenian>, dostop 12. 8. 2026.
- [20] J. Kim, J. Kong, J. Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. ICML, 2021. arXiv:2106.06103.
- [21] S. Mehta in sod. Matcha-TTS: A fast TTS architecture with conditional flow matching. ICASSP, 2024. arXiv:2309.03199. Koda MIT.
- [22] Y. Chen in sod. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. arXiv:2410.06885, 2024. Koda MIT; objavljene uteži CC BY-NC 4.0.
- [23] S. E. Eskimez in sod. E2 TTS: Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS. IEEE SLT, 2024. arXiv:2406.18009.
- [24] F. Lux in sod. Meta Learning Text-to-Speech Synthesis in over 7000 Languages. Interspeech, 2024. arXiv:2406.06403. Ogrodje IMS-Toucan in model ToucanTTS: <https://github.com/DigitalPhonetics/IMS-Toucan>;
<https://huggingface.co/Flux9665/ToucanTTS> (Apache 2.0).
- [25] Y. A. Li in sod. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. NeurIPS, 2023. arXiv:2306.07691.
- [26] E. Casanova in sod. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. Interspeech, 2024. arXiv:2406.04904. Uteži pod licenco CPML (nekomercialna).
- [27] Z. Du in sod. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. arXiv:2412.10117, 2024; CosyVoice 3. arXiv:2505.17589, 2025. Apache 2.0.
- [28] Kokoro-82M. <https://huggingface.co/hexgrad/Kokoro-82M> (Apache 2.0; 9 jezikov), dostop 12. 8. 2026.
- [29] V. Pratap in sod. Scaling Speech Technology to 1,000+ Languages (MMS). arXiv:2305.13516, 2023. Koda in uteži CC BY-NC 4.0 (fairseq/examples/mms).
- [30] Resolucija o nacionalnem programu za jezikovno politiko 2021-2025 (ReNPJP21-25). Uradni list RS, 2021.
<https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2021-01-1999/>.
- [31] Nacionalna strategija za razvoj umetne inteligence do leta 2030 (NsUI 2030). Ministrstvo za digitalno preobrazbo.
<https://www.gov.si/assets/ministrstva/MDP/DDD/NsUI2030.pdf>.
- [32] ALT-EDIC: Alliance for Language Technologies, European Digital Infrastructure Consortium.
<https://www.alt-edic.eu/>, dostop 16. 8. 2026.
- [33] Morfološki leksikon Sloleks 3.1. CLARIN.SI. <http://hdl.handle.net/11356/2080> (CC BY-SA 4.0; februar 2026).

- [34] Automatically stress labelled morphological lexicon Sloleks 1.2. CLARIN.SI. <http://hdl.handle.net/11356/1186> (CC BY-NC-SA 4.0).
- [35] J. Žganec Gros in sod. SI-PRON pronunciation lexicon: a new language resource for Slovenian. LREC, 2006. <https://aclanthology.org/L06-1057/>. Vir javno ni dostopen.
- [36] V. Smole. Slovenska narečja po klasifikaciji Logar-Rigler. 55. SSJLK, Univerza v Ljubljani, 2019. https://centerslo.si/wp-content/uploads/2019/06/55-SSJLK_Smole.pdf; Karta slovenskih narečij: https://sla.zrc-sazu.si/eSLA/Karta_narecij_600dpi_CMYK.pdf.
- [37] Interaktivna karta slovenskih narečnih besedil. <https://narecja.si/>; opis sistema: zbornik JTDH, 2018. https://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Kavcic-et-al_Interaktivna-karta-slovenskih-narecnih-besedil.pdf.
- [38] Franček, narečni modul (ZRC SAZU). CLARIN.SI. <http://hdl.handle.net/11356/1473> (CC BY 4.0).
- [39] DIALECT-COPA (vključno s cerkljanskim narečjem). CLARIN.SI. <http://hdl.handle.net/11356/1766> (CC BY-SA 4.0).
- [40] Govorni korpus Gos 2.1. CLARIN.SI: transkripcije <http://hdl.handle.net/11356/1863> (CC BY-SA 4.0); posnetki <http://hdl.handle.net/11356/1973> (omejena licenca ACA ID-BY-INF-NORED).
- [41] Učni korpus govorne slovenščine ROG 1.1. CLARIN.SI. <http://hdl.handle.net/11356/2062> (CC BY-SA 4.0).
- [42] Mozilla Common Voice, slovenščina, izdaja 26.0 (junij 2026): uradne statistike. <https://github.com/common-voice/cv-dataset/blob/main/datasets/scripted-speech/cv-corpus-26.0-2026-06-12.json> (CC0).
- [43] VoxPopuli (Meta AI). <https://github.com/facebookresearch/voxpopuli> (CC0; slovenščina: 4.400 ur v1 oziroma 11.300 ur v2 neoznačenih; 10 ur transkribiranih).
- [44] M. Gaido in sod. MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages. EMNLP, 2024. arXiv:2410.01036. Podatki: <https://huggingface.co/datasets/GBK-MT/mosel> (CC BY 4.0; slovenščina: 5.851 ur samodejnih prepisov).
- [45] YODAS (espnet). <https://huggingface.co/datasets/espnet/yodas> (CC BY 3.0; slovenščina sl000: 22,9 h); YODAS2: <https://huggingface.co/datasets/espnet/yodas2>.
- [46] A. Conneau in sod. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. arXiv:2205.12446, 2022. <https://huggingface.co/datasets/google/fleurs> (CC BY 4.0).
- [47] M. Ma in sod. FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks. arXiv:2408.06227, 2024. <https://huggingface.co/datasets/google/fleurs-r> (CC BY 4.0).
- [48] OpenSLR, seznam virov (brez slovenščine). <https://openslr.org/resources.php>, dostop 16. 8. 2026.
- [49] Korpus Gigafida 2.0. CLARIN.SI. <http://hdl.handle.net/11356/1320>; ccGigafida: <http://hdl.handle.net/11356/1035> (CC BY-NC-SA).
- [50] espeak-ng: seznam podprtih jezikov (slovenščina, koda sl). <https://github.com/espeak-ng/espeak-ng/blob/master/docs/languages.md> (GPLv3).
- [51] Orodja RSDO: normalizator <http://hdl.handle.net/11356/1742>; denormalizator <http://hdl.handle.net/11356/1743>; vstavljalnik ločil <http://hdl.handle.net/11356/1735> (CLARIN.SI).
- [52] CLASSLA (CLARIN.SI): cevovod za obdelavo slovenščine <https://github.com/clarinsi/classla> (Apache 2.0); razvrščevalnik primarnega naglasa <https://huggingface.co/classla/Wav2Vec2BertPrimaryStressAudioFrameClassifier>.
- [53] J. Križaj, J. Žganec Gros, S. Dobrišek. Prisilna poravnava za slovenščino na osnovi MFA (akustični model in slovar). JTDH, 2024. https://github.com/jan3zk/forced_alignment (Apache 2.0).
- [54] F. Lux, N. T. Vu. Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features. ACL, 2022. <https://aclanthology.org/2022.acl-long.472/>.
- [55] E. Casanova in sod. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. ICML, 2022. arXiv:2112.02418.
- [56] E. J. Hu in sod. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685, 2021.
- [57] SLAIF, slovenska tovarna umetne inteligence: načini dostopa. <https://www.slaif.si/>; https://www.eurohpc-ju.europa.eu/ai-factories/ai-factories-access-modes_en, dostop 12. 8. 2026.
- [58] ITU-T P.800: Methods for subjective determination of transmission quality, 1996; ITU-T P.808: Subjective evaluation of speech quality with a crowdsourcing approach, 2021.
- [59] T. Saeki in sod. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. Interspeech, 2022. arXiv:2204.02152.
- [60] G. Mittag in sod. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. Interspeech, 2021. arXiv:2104.09494.
- [61] SloBench: evalvacijska platforma za slovenske jezikovne tehnologije. <https://slobench.cjvt.si/>, dostop 17. 8. 2026.