

Informacijska tabela projekta / Project information card	
Akronim projekta / Project acronym	LLM4DH
Naslov projekta / Project title	Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities
Spletna stran / Website	https://www.cjvt.si/llm4dh/
Št projekta / Project No.	GC-0002
Razpis / Call	Javni razpis za (so)financiranje Gravitacije, št: 5100-70/2024-6 z dne 7.5.2024 in Sprememb javnega razpisa za (so)financiranje Gravitacije, št. 5100-70/2024-11 z dne 29.5.2024.
Koodrinator projekta / Project coordinator	Marko Robnik-Šikonja
Trajanje projekta / Project duration	1.10.2024 – 30.9.2027, 36 mesecev / 36 months
Povzetek projekta / Abstract of the project	Jezikovne tehnologije igrajo vse pomembnejšo vlogo na številnih področjih znanosti. Z bliskovitim napredkom velikih jezikovnih modelov se ti pozicionirajo kot osrednja tehnologija umetne inteligence z daljnosežnim vplivom ne le na znanstveno raziskovanje ampak tudi na širši družbeni razvoj in celotno družbo. Veliki jezikovni modeli, kljub svojim velikim zmožnostim, izkazujejo vrsto pomanjkljivosti, kot so nekonsistentni in napačni odgovori, velike računske zahteve, šibko poznavanje jezikov z manj viri, neprilagojenost za nekatere pomembne naloge in domene ter slabosti v razumevanju družbe, etike in človekovih potreb. Projekt bo tehnologijo velikih jezikovnih modelov izboljšal z različnimi znanji in podatki, jim zmanjšal računsko zahtevnost pri govornih tehnologijah in jim omogočil dostop do zanesljivih zunanjih informacij. Izboljšave v temeljnih zmožnostih velikih jezikovnih modelov bodo vodile do prebojnih dosežkov na več področjih digitalne humanistike: v jezikoslovju, leksikografiji, izobraževanju, analizi družbenih pojavov in sodobne zgodovine, političnih vedah, folkloristiki in pravu. Language technologies are increasingly important in many areas of science. With the rapid advances in large-scale language models, they are positioning themselves as a core AI technology with far-reaching impact on scientific research, wider societal development, and society. Despite their great capabilities, large language models exhibit several shortcomings, such as inconsistent and incorrect answers, high computational demands, weak knowledge of less-resourced languages, poor performance in some important tasks and domains, and weaknesses in understanding society, ethics, and human needs. The project will improve the technology of large language models by enriching them with diverse knowledge and data, using them to improve the performance of speech technologies, and giving them access to reliable external information. Improvements in the core capabilities of large language models will lead to breakthroughs in several areas of the digital humanities: linguistics, lexicography, education,

	analysis of social phenomena and contemporary history, political sciences, folkloristics, and law.
Vodilni partner prprojekta/ Lead project partner	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko (UL FRI):
Sodelujoči partnerji / Participating project partners	Institut "Jožef Stefan" (IJS) Inštitut za novejšo zgodovino (INZ) Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja (IRRIS) Inštitut za kriminologijo pri Pravni Fakulteti v Ljubljani (IK) Univerza v Ljubljani, Filozofska fakulteta (UL FF) Univerza v Ljubljani, Fakulteta za elektrotehniko (UL FE) Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko (UM FERI)

Zgodovina dokumenta			
Datum	Verzija	Avtor	Komentar
1. 9. 2026	1.0	Maja Zupančič Justin	Postavitev zasnove poročila
1. 10. 2026	1.1	Vsi vodje DS in DN	dopolnitve
8.10.2026	1.2	Marko Robnik Šikonja	Zaključna redakcija

Kazalo vsebine / Table of content

POVZETEK REZULTATOV IN DOSEŽKOV PROJEKTA V DRUGEM LETU IZVAJANJA	5
1. REZULTATI IN DOSEŽKI PROJEKTA: REALIZACIJA PROGRAMA DELA V DRUGIH DVANAJSTIH MESECIH TRAJANJA PROJEKTA OD M13 DO M24	8
1.1. IZZIV 1: IZBOLJŠANJE METOD VELIKIH JEZIKOVNIH MODELOV Z JEZIKOVNIMI VIRI IN RAZVOJ SLIKOVNO-JEZIKOVNIH MODELOV	8
1.1.1. Aktivnost 1.1: Izboljševanje velikih jezikovnih modelov z jezikovnimi podatki (M1-M24)	8
1.1.2. Aktivnost 1.2: Izboljšanje velikega jezikovnega modela z medjezikovnimi podatki (M13-M30)....	9
1.1.3. Aktivnost 1.3: Izboljšanje multimodalnih modelov (M1-M24).....	9
1.2. IZZIV 2: VELIKI JEZIKOVNI MODELI ZA JEZIKOSLOVJE IN UPRAVLJANJE ZNANJA	10
1.2.1. Aktivnost 2.1: Veliki jezikovni modeli za učinkovito leksikografijo (M1-M36)	10
1.2.2. Aktivnost 2.2: Nevronsko preverjanje pravopisa in slovnice (M1-M36).....	10
1.2.3. Aktivnost 2.3: Napredna slovnična analiza večjezičnih korpusov (M1-M36)	11
1.3. IZZIV 3: VELIKI JEZIKOVNI MODELI ZA GOVORJENI JEZIK.....	12
1.3.1. Aktivnost 3.1: Učinkovito zbiranje podatkov o govorjenem jeziku (M1-M36)	12
1.3.2. Aktivnost 3.2: Semantično in pragmatično procesiranje govora (M7-M30).....	13
1.3.3. Aktivnost 3.3: Nadzor kakovosti in filtriranje govornih podatkov (M1-M36).....	14
1.3.4. Aktivnost 3.4: Veliki jezikovni modeli za prepoznavo domensko specifičnega govora (M1-M36).....	14
1.4. IZZIV 4: NAPREDNE TEHNOLOGIJE ZA DIGITALNO HUMANISTIKO	15
1.4.1. Aktivnost 4.1: Grafi imenskih entitet (M1-M18)	15
1.4.2. Aktivnost 4.2: Veliki jezikovni modeli za diahrono analizo (M6-M24).....	16
1.4.3. Aktivnost 4.3: Multimodalni modeli za analizo slik (M10-M30)	16
1.4.4. Aktivnost 4.4: Veliki jezikovni modeli in RAG za detekcijo kontradikcij (M1-M18)	17
1.5. IZZIV 5: IZBRANI IZZIVI DIGITALNE HUMANSTIKE	18
1.5.1. Aktivnost 5.1: Veliki jezikovni modeli za historiografijo (M1-M36).....	18
1.5.2. Aktivnost 5.2: Veliki jezikovni modeli za folkloristiko (M1-M36)	18
1.5.3. Aktivnost 5.3: Veliki jezikovni modeli in RAG za pravno domeno (M1-M36)	20
1.6. IZZIV 6: EVALVACIJA IN RAZUMEVANJE VELIKIH JEZIKOVNIH MODELOV	20
1.6.1. Aktivnost 6.1: Figurativni jezik in pragmatični benčmarki (M1-M24).....	21
1.6.2. Aktivnost 6.2: Primerjalni test razumevanja govorjenega jezika (M25-M36).....	21
1.6.3. Aktivnost 6.3: Detekcija pristranskosti v velikih jezikovnih modelih in pri avtomatski prepoznavi govora (M10-M36).....	22
1.6.4. Aktivnost 6.4: Razlage velikih jezikovnih modelov na osnovi znanja (M19-M36).....	23
1.7. DELOVNI SKLOP 7: UPRAVLJANJE IN DISEMINACIJA	23
1.7.1. Aktivnost 7.1: Koordinacija in spremljanje projekta (M1-M36)	23
1.7.2. Aktivnost 7.2: Upravljanje etike, kakovosti, tveganj in podatkov (M1-M36)	24
1.7.3. Aktivnost 7.3: Komunikacija, diseminacija in eksploatacija (M1-M36).....	24
2. NAJPOMEMBNEJŠI DOSEŽKI PROJEKTNE SKUPINE V DRUGIH DVANAJSTIH MESECIH TRAJANJA PROJEKTA	25
2.1. DOSEŽKI NA RAZISKOVALNEM PODROČJU	25
2.1.1. Izziv 1 Izboljšanje metod velikih jezikovnih modelov z jezikovnimi viri in razvoj slikovno-jezikovnih modelov	25
2.1.2. Izziv 2 Veliki jezikovni modeli za jezikoslovje in upravljanje znanja	26
2.1.3. Izziv 3 Veliki jezikovni modeli za govorjeni jezik	26
2.1.4. Izziv 4 Napredne tehnologije za digitalno humanistiko.....	27
2.1.5. Izziv 5 Izbrani izzivi digitalne humanstike.....	29
2.1.6. Izziv 6 Evalvacija in razumevanje velikih jezikovnih modelov.....	29
2.2. DRUGI POMEMBNI REZULTATI	31

Povzetek rezultatov in dosežkov projekta v drugem letu izvajanja

Projekt LLM4DH (ARIS, GC-0002) je drugo leto izvajanja (M13–M24) zaključil brez odstopanj od delovnega načrta in s preseženimi vsemi kazalniki. Konzorcij osmih partnerjev pod vodstvom UL FRI je objavil prve slovenske slikovno-jezikovne modele, izboljšane leksikografske modele GaMS, vrsto odprtih podatkovnih zbirk in primerjalnih testov ter 15 znanstvenih člankov v revijah ter 30 konferenčnih prispevkov. Rezultati so prosto dostopni na platformah Hugging Face, CLARIN.SI in GitHub, številni pa presegajo slovenščino in služijo raziskavam več sto svetovnih jezikov.

Ključni dosežki drugega leta

- **Bistveno izboljšana Digitalna slovarska baza.** Z velikimi jezikovnimi modeli smo generirali paradigme, definicije, kolokacije, oznake in opise figurativnega jezika; baza je odprto dostopna kot zbirka in prek API-ja.
- **Leksikografski veliki jezikovni modeli.** Model GaMS3-12B-Instruct-Lex odgovarja na vprašanja o slovarskih podatkih in je povezan s cevovodom RAG nad Digitalno slovarsko bazo (DSB). Prva večjezična različica vključuje 77.000 prevodov pomenov v madžarščino.
- **Prvi slovenski slikovno-jezikovni modeli.** Modela SVILA-1-4B in SVILA-1-12B, naučena na več kot milijonu parov slika–besedilo, bistveno presegata izhodiščne modele pri vizualnem odgovarjanju na vprašanja, OCR in povezovanju opisov z deli slik. Modelu GaMS3-12B smo dali polne multimodalne zmožnosti.
- **Slovnično popravljajanje na ravni vrhunskih modelov.** Prilagojen model GaMS2-27B-GEC dosega natančnost 0,657 in priklic 0,648, kar je enako ali bolje od modela Claude-Opus-4.8 in močno presega Gemini-3-flash. Pri številnih oblikoslovnih problemih presega mejo $F_{0.5} = 0,7$ za aplikativno rabo. Zbirka DASSLE 2.0 obsega 300.000 poravnanih povedi.
- **Govorni viri.** Občani so prek portala Govorjena slovenščina prispevali okoli 50 ur pogovorov iz vseh statističnih regij. Objavljena sta korpus ROG-dialog 1.0 z dialoškim dejanji in sentimentom ter korpus Zila s 100 urami ziljskega narečja.
- **Nov pristop k vrednotenju razpoznavе govora.** Novo večreferenčno vrednotenje upošteva, da spontani govor nima enega samega pravilnega prepisa; izmerjeno napako zmanjša do 20 % in bolje korelira s semantičnimi ocenami.
- **Pravna domena.** Korpus COLESLAW (540.000 dokumentov, 770 milijonov besed) je največji korpus slovenskega pravnega jezika. Sistem ContrAG dokaj zanesljivo zaznava neskladne člene zakonov v manjših zbirkah ($\text{acc}@10 = 0,976$).
- **Digitalna humanistika v praksi.** Napovedi o odnosu do kolektivnih identitet smo razširili na 2,65 milijona omemb v zgodovinskem časopisju; orodje Semantic Shift Viewer omogoča sledenje pomenskim premikom v zgodovinskih korpusih; večjezični korpus 108 del (16,9 milijona besed) odpira primerjalne raziskave reševanja konfliktov.

Projekt se z rezultati uveljavlja kot referenčni primer, kako jezik z manj viri razvije lastne, odprte in konkurenčne velike jezikovne modele ter kaže kako lahko raziskave, ki združujejo umetno inteligenco in digitalno humanistiko, pripomorejo k razvoju obeh področij ter napredku širše družbe.

Rezultati projekta dosegajo svetovno pomembnost na več ravneh.

- **Viri za svetovne jezike.** WALB-Bench preverja metajezikovno znanje modelov za 192 jezikovnih značilnosti v več kot 2.600 jezikih; GramQA in agentni sistem UD-Agent pokrivata 179 jezikov iz zbirke Universal Dependencies. Rezultati so sprejeti v reviji *Natural Language Processing*.
- **Regionalni in evropski viri.** PragMega smo lokalizirali v češčino, slovaščino in hrvaščino; ParlaCAP zajema 28 evropskih parlamentov; CLASSLA-web 2.0 in ParlaSpeech 3.0 krepi vire

za južnoslovanske in srednjeevropske jezike; pripravili smo taksonomijo in register evropskih primerjalnih testov za velike jezikovne modele.

- **Metode, prenosljive na druge jezike.** Večreferenčno vrednotenje razpoznavе govora, delotok sintetičnih jezikovnih napak, cevovod RAG nad slovarskimi podatki in metodologija občanske znanosti za zbiranje govora so neposredno uporabni za druge jezike z manj viri.
- **Objave v vrhunskih revijah.** Med 15 članki v revijah so objave v revijah *Computer Speech & Language*, *Machine Learning*, *Journal of Machine Learning Research*, *International Journal on Document Analysis and Recognition*, *Language and Speech in Frontiers in Artificial Intelligence*. Več prispevkov je bilo predstavljenih na konferenci LREC 2026 in njenih delavnicah, prispevek na konferenci ACL pa je prejel priznanje »častna omemba«.
- **Mednarodna skupnost.** Projekt je organiziral mednarodno multikonferenco SyntaxFest 2025 in soorganiziral konferenco JT-DH 2026; rezultati so bili predstavljeni na konferencah CLARIN 2025 na Dunaju, SPECOM, IEEE ELEKTRO in delavnici LUHME ob EMNLP 2026.
- **Mreženje in sodelovanje.** Raziskovalci sodelujejo v akciji COST CHANGECODE, evropskem centru odličnosti AI4DH, bilateralnem projektu z ZDA ter s podjetjem NVIDIA v pobudi Sovereign AI.

Delo v vseh šestih raziskovalnih izzivih poteka po načrtu, obljubljeni izročki so javno objavljeni.

Izziv	Glavni rezultati	Javno dostopni izdelki
1 Izboljšani modeli in slikovno-jezikovni modeli	Leksikografski GaMS z RAG nad DSB, prva večjezična različica, slovenski slikovno-jezikovni modeli, kulturno-semiotična zbirka (50.000 primerov)	GaMS3-12B-Instruct-Lex, GaMS3-12B-Multimodal, SVILA-1-4B, SVILA-1-12B
2 Jezikoslovje in upravljanje znanja	Izboljšana DSB, model za slovnično popravljanje GaMS2-27B-GEC, avtentične evalvacijske množice, agentna korpusna slovnična analiza	DASSLE 2.0, WALS-Bench 1.0, GramQA 1.0, UD-Agent
3 Govorjeni jezik	Občanska znanost (≈50 ur novih posnetkov), anotacija dialoških dejanj in sentimenta, izhodiščni razpoznavnik govora, študija povezovanja razpoznavе govora z modeli	ROG-dialog 1.0, korpus Zila in narečni modeli
4 Tehnologije za digitalno humanistiko	Grafi imenskih entitet z vizualizacijo negotovosti, diahrona analiza pomena, multimodalna zbirka zgodovinskega časopisja, ContRAG za pravna neskladja	Zbirka slik zgodovinskega časopisja, SLawNLI
5 Izbrani izzivi digitalne humanistike	Analize ideoloških konceptov v zgodovinskem tisku, metodologija računske folkloristike, korpus Ciciban (≈10.000 strani), pravni klepetalnik in sledenje zakonodaji za podjetja	COLESLAW, Semantic Shift Viewer, označena zbirka 371 omemb
6 Evalvacija in razumevanje modelov	Primerjalni testi pragmatičnega in asociativnega vedenja, SloBENCH2 z večreferenčnim vrednotenjem, analize pristranskosti (spol, etničnost, politika, morala), razlage na osnovi znanja	PragMega CzeSloCro, WAX-SL, EEC-SL

Izziv	Glavni rezultati	Javno dostopni izdelki
7 Upravljanje in diseminacija	Delo po načrtu, štiri polletna srečanja konzorcija, etični kodeks in načrt ravnanja s podatki	Letno poročilo M24

Diseminacija je močno presegla cilje: 15 člankov (cilj 7), 30 konferenčnih prispevkov (cilj 7), 90 objav na družbenih omrežjih (cilj 1 na teden) in 16 medijskih objav na RTV Slovenija in POP TV (cilj 5). Na razpisu ARIS je bil uspešen tudi nadgrajevalni projekt *Od občanske znanosti do digitalne slovarske baze*.

Pogled naprej

V zadnjem letu bo projekt izboljšano DSB uporabil za učenje velikih modelov, razširil večjezični leksikografski model na 27 jezikov ter razvil slikovno-jezikovni model za digitalno humanistiko. Na portal SloBench bodo uvrščeni primerjalni testi razumevanja govorjenega jezika in slovničnega popravljanja, nadgrajena bosta UD-Agent in ContrAG, nastala pa bo tudi govorna zbirka z najmanj 300 urami posnetkov. S tem bo LLM4DH ustvaril celovit, odprt in mednarodno primerljiv nabor jezikovnih tehnologij za digitalno humanistiko in slovenščino ter ustvaril vrsto mednarodno pomembnih znanstvenih rezultatov.

1. Rezultati in dosežki projekta: realizacija programa dela v drugih dvanajstih mesecih trajanja projekta od M13 do M24

1.1. Izziv 1: Izboljšanje metod velikih jezikovnih modelov z jezikovnimi viri in razvoj slikovno-jezikovnih modelov

Veliki jezikovni modeli za učenje in prilagajanje določenim nalogam potrebujejo velike količine visokokakovostnih besedilnih podatkov. Takšni podatki so lahko v pomoč pri predhodnem učenju velikih jezikovnih modelov, saj je z njimi možno pripraviti različne vrste podatkov, zlasti grafe znanja in neobdelano besedilo. Informacije, ki so na voljo v leksikografskih virih, vključujejo relacije, informacije o porazdelitvi pomenov z definicijami besednih pomenov, medjezikovne povezave, identifikacijo in opis idiomatskih besednih zvez ali frazemov itd.

Uporabo te zakladnice informacij, ki je veliki jezikovni modeli še ne uporabljajo ustrezno, smo raziskovali v drugem letu projekta v nalogah T1.1 in T1.2. V nalogi T1.3 pa smo se osredotočili na podporo slikovnemu gradivu, predvsem slovenskemu, ki ga slovenski veliki jezikovni modeli predtem še niso znali ustrezno obravnavati.

1.1.1. Aktivnost 1.1: Izboljševanje velikih jezikovnih modelov z jezikovnimi podatki (M1-M24)

V aktivnosti T1.1 smo do M12 izdelali začetno verzijo velikega slovenskega jezikovnega modela GaMS, podprtega z informacijami iz digitalne slovarske baze (DSB). V času od M13 do M24 se osredotočili na odpravo pomanjkljivosti začetnega modela, ki je nastal v prvem letu, in bistveno izboljšanje kakovosti učnih podatkov tako za fazo nadaljevanja vnaprejšnjega učenja (*continual pretraining, CPT*) kot za fazo prilagajanja odgovarjanju na vprašanja. V učni množici za vnaprejšnje učenje modela smo opise besed prilagodili tako, da so oblikovani bolj podobno naravnemu besedilu, ki bi ga pričakovali v odgovorih modela in so uporabnikom bolj prijazni. Pri pripravi podatkov za odgovarjanje na vprašanja smo šli skozi več iteracij podatkovne zbirke. Eksperimentirali smo z uporabo velikega jezikovnega modela za pripravo parov vprašanj in odgovorov ter se na koncu odločili za pripravo zbirke s strogo določenimi oblikami odgovorov. S tem smo omejili težave pri kvaliteti odgovorov zaradi manjkajočih pregibnih oblik ter vdorov tujih jezikov. Pripravili smo tudi zbirko vprašanj in odgovorov o splošnih lastnostih slovenščine, s katero smo nekoliko izboljšali sposobnost modela pri odgovarjanju na bolj splošna vprašanja.

Poleg izboljšav modela smo za odgovarjanje na leksikografska vprašanja pripravili tudi cevovod RAG (angl. Retrieval Augmented Generation, s pridobivanjem obogateno generiranje), ki omogoča iskanje odgovora v DSB. Za cevovod smo sprva pripravili obsežen graf znanja v OntoLex-Lemon shemi, v katerega smo uvozili ključne podatke iz DSB. Kasneje smo graf znanja pretvorili v relacijsko bazo, po kateri lahko iščemo z uporabo SQL izrazov, saj se je izkazalo, da obstoječi veliki jezikovni modeli bolje iščejo po relacijskih kot po grafovskih bazah podatkov. Med razvojem cevovoda smo prilagodili tudi leksikografski veliki jezikovni model za iskanje po grafu znanja, vendar se je kljub temu hranjenje podatkov v relacijski bazi izkazalo kot bolj učinkovito.

Nazadnje smo izvedli tudi nekaj dodatnih eksperimentov z uporabo novega grafovsko-jezikovnega modela osnovanega na modelu GaMS3-12B, ki združuje jezikovni vhod z vhodnim grafom znanja. Pri tem so prvi eksperimenti pokazali obetavne rezultate, ki nakazujejo na to, da je takšen model sposoben iz konteksta v obliki grafa znanja dobro izvleči informacije, ključne za odgovor na uporabnikovo vprašanje.

Kazalnik 1.1:

- [Končni izboljšani veliki jezikovni modeli](#) (M24, na voljo).

1.1.2. Aktivnost 1.2: Izboljšanje velikega jezikovnega modela z medjezikovnimi podatki (M13-M30)

V okviru izboljševanja modela z medjezikovnimi podatki smo se v tej začetni fazi naloge osredotočili na testno vključitev enega jezika, in sicer madžarskega, ki je že del digitalne slovarske baze (DSB). Iz baze smo pridobili in v sistem integrirali približno 77.000 prevodov pomenov ter 18.000 prevedenih zgledov rabe, medtem ko pregibne oblike in definicije za ta jezik še niso zajete v DSB. S tem smo modelu omogočili odgovarjanje na poizvedbe o dvojezičnih prevodih posameznih besed (npr. iskanje prevodov za določeno besedo ali prikaz prevedenih primerov) ter na splošnejša vprašanja o madžarskih frazah in jezikovnih posebnostih.

S podatki o madžarskem jeziku smo doučili leksikografski veliki jezikovni model osnovan na modelu Gemma 3. V naslednjih korakih načrtujemo pretvorbo podatkovnih zbirk DBnary (ki v formatu OntoLex-Lemon pokriva 27 jezikov z leksikalnimi enotami, oblikami, pomeni in definicijami) ter dvojezičnih slovarjev DMLex v našo kompaktno relacijsko shemo SQL za uporabo v RAG cevovodu. Končni cilj je preslikava in uvoz teh večjezičnih virov v DSB, kar bo omogočilo zanesljiv medjezikovni prenos in poenoteno obravnavo leksikografskih vprašanj v več jezikih.

Dodatno smo v tej nalogi začeli preučevati možnost za izboljšavo velikih jezikovnih modelov za strojno prevajanje figurativnega jezika, kjer kot vir informacij uporabljamo večjezične slovarje.

Kazalnik 1.2:

- [Graf znanja in podatkovne zbirke neobdelanih besedil](#) (M18, na voljo).
- [Začetni izboljšani veliki jezikovni model](#) (M24, na voljo).

1.1.3. Aktivnost 1.3: Izboljšanje multimodalnih modelov (M1-M24)

V okviru projekta sta bila zastavljena dva ključna cilja: vzpostavitev obsežne slovenske podatkovne zbirke za učenje velikega slikovno-jezikovnega modela (M12) in kasnejši razvoj slovenskega slikovno-jezikovnega modela (M24).

V prvem letu projekta smo uspešno dosegli prvi ključni mejnik (M12) z objavo obsežne podatkovne zbirke za učenje slikovno-jezikovnih modelov za slovenski jezik, poimenovano SLO-VLM-IT-Dataset 1.0. Ta zbirka je postala temelj za drugi ključni mejnik (M24), ki smo ga uspešno dosegli z razvojem, celovitim vrednotenjem in javno objavo prvih namenskih slovenskih slikovno-jezikovnih modelov (SJM) ter uspešno diseminacijo dosežkov v mednarodni znanstveni skupnosti. Celotno metodologijo priprave obsežne multimodalne zbirke in postopek prilagajanja modelov smo predstavili v znanstvenem članku na ugledni mednarodni konferenci LREC.

Z učenjem na zgrajeni podatkovni zbirki, ki vsebuje več kot milijon parov slika-besedilo, smo z metodo nadzorovanega prilagajanja (angl. *instruction fine-tuning*) na osnovi večjezičnih temeljnih modelov Gemma-3 razvili dva nova modela, poimenovana SVILA (Slovenian Vision Language Assistant). Gre za različici SVILA-1-4B s štirimi milijardami parametrov ter večjo in zmogljivejšo SVILA-1-12B z dvanajstimi milijardami parametrov. Oba modela sta skupaj z vso izvirno kodo za učenje prosto dostopna za raziskovalno in aplikativno rabo na platformi Hugging Face, koda za učenje pa je dostopna v repozitoriju GitHub. Kot pomemben komplementaren tehnični dosežek smo z zgrajeno zbirko podatkov izvedli še ponovno poravnavo besedilnega kodirnika modela GaMS-3-12B s slikovnim kodirnikom (angl. *vision encoder*). S tem smo modelu GaMS-3-12B povrnili polne multimodalne zmožnosti, nov multimodalni model pa je dostopen na platformi [Hugging Face](#).

Za preverjanje zmogljivosti modelov smo naredili novo, strojno prevedeno in ročno pregledano slovensko evalvacijsko zbirko. Rezultati vrednotenja so potrdili učinkovitost pristopa, saj modela SVILA

dosegata bistveno boljše rezultate v primerjavi z izhodiščnimi modeli pri nalogah vizualnega odgovaranja na vprašanja (angl. visual question answering, VQA), optične prepoznavne znakov (angl. optical character recognition, OCR) ter povezovanja opisov z deli slik (angl. *visual grounding*). Slabost te evalvacijske množice je, da gre za prevod iz angleščine, ki ne pokriva vseh vidikov slovenske kulture in življenja v Sloveniji, zato smo pred kratkim razvili še [prvo slovensko kulturno-semiotično podatkovno zbirko](#), namenjeno učenju in evalvaciji višjih ravni vizualne interpretacije. Nova zbirka, ki obsega okoli 50 tisoč učnih primerov, omogoča učenje in evalvacijo SJM glede prepoznavanja slovenskih kulturno pogojenih, simbolnih in konotativnih pomenov.

Kazalniki 1.3:

- Slovenska slikovno-jezikovna modela [SVILA-1-4B](#) in [SVILA-1-12B](#) (M24, na voljo).

1.2. Izziv 2: Veliki jezikovni modeli za jezikoslovje in upravljanje znanja

Slovenščina s svojo dvomilijonsko jezikovno skupnostjo je primer jezika z manj viri, kar je pokazala tudi primerjava, ki jo je pripravil projekt European Language Equality (Rehm & Way, 2023). Raziskovalni izziv te naloge je omogočiti izdelavo kakovostnih semantičnih opisov za jezike z manj viri in omogočiti obsežne jezikovne primerjave, ki bodo omogočile nov vpogled v slovnico svetovnih jezikov.

1.2.1. Aktivnost 2.1: Veliki jezikovni modeli za učinkovito leksikografijo (M1-M36)

Nadaljevali smo z razvojem različnih avtomatiziranih in z velikimi jezikovnimi modeli podprtih postopkov za analizo jezikovnih podatkov in izdelavo leksikografskih opisov. Digitalno slovarsko bazo (DSB) za slovenščino izboljšujemo na različnih ravneh jezikovnega opisa z uporabo tako komercialnih velikih jezikovnih modelov (VJM), modelov izdelanih v aktivnosti T1.1 in drugih odprtodostopnih modelov. Generiramo morfološke in semantične podatke, pri čemer smo se osredotočili na: 1) oblikovanje morfoloških paradigem, 2) besednovrstno razločevanje, 3) oblikovanje različnih vrst definicij (pomenski kazalniki, poenostavljene, terminološke itd.), 4) izboljšanje kolokacij in primerov rabe; 5) pripisovanje oznak (slogovne, normativne, področne, žanrske itd.) 6) opis idiomatskega, figurativnega in metaforičnega jezika itd. Pripravljena je prva različica bistveno izboljšane DSB, ki jo bomo v naslednjem letu uporabili za izboljšanje modelov v aktivnosti T1.1. Različice DSB so na voljo kot javno dostopne podatkovne zbirke in prek API-ja z odprtim dostopom.

Kazalnik 2.1:

- Digitalna slovarska baza z generiranimi leksikografskimi podatki – prva različica (M24, na voljo)

1.2.2. Aktivnost 2.2: Nevronske preverjanje pravopisa in slovnice (M1-M36)

Cilja naloge T2.2 v poročanem obdobju sta bila na podlagi podatkovne množice sintetičnih jezikovnih napak (M12) izdelati veliki jezikovni modeli za preverjanje slovnice (M18) ter izdelati novo podatkovno množico z avtentičnimi podatki za evalvacijo strojnega slovničnega pregledovanja (M24).

V drugem letu projekta smo razvili napreden delotok priprave povedi s sintetičnimi jezikovnimi napakami. Izhajali smo iz zbirke *DASSLE 1.0*, ki smo jo pripravili prejšnje leto, a so prva učenja modelov pokazala potrebo po večji količini gradiva za posamezen jezikovni problem. Novi delotok obsega: natančne jezikoslovne analize za posamezen jezikovni problem; pripravo ciljnega gradiva z označenim jezikovnim elementom, ki ga je mogoče strojno "pokvariti" na določen način; oblikovanje ciljnih

pozivov, ki vsebujejo vrsto prilagojenih varovalk za kontrolo kvalitete, vključno s preverjanjem, ali so strojno "pokvarjene" povedi dejansko jezikovno nesprejemljive ali ne. Rezultat dela je 300.000 poravnanih povedi (ena z jezikovnim problemom in druga z jezikovnim popravkom) za 20 specifičnih oblikoslovnih težav. Podatki so dostopni na platformi [HuggingFace](#). Novi delotok smo predstavili v dveh konferenčnih prispevkih (Štebljaj idr. 2026, Arhar Holdt idr. 2026).

Na manjši količini gradiva smo testirali štiri modele s predvideno visoko ravni znanja za slovenščino (Gemma-3-27B-IT, GaMS3-12B, GaMS3-12B-Instruct, GaMS2-27B). Najbolje se je obnesel GaMS2-27B, ki smo ga nato nadalje doučili v dveh korakih: najprej na omenjenih sintetično pripravljenih podatkih, nato pa še na avtentičnih podatkih iz korpusa KOST (13.000 povedi). V učenje smo vključili tudi 20 % povedi brez jezikovnega problema, saj je ključno naučiti popravjalnike, da jezikovno ustreznih povedi ne spreminjajo. Na množici DASSLE 1.0 smo evalvirali doučeni model GaMS2-27B-GEC (natančnost: 0.657, priklic: 0.648) primerjalno z modeloma Claude-Opus-4.8 (natančnost: 0.657; priklic: 0.585) in Gemini-3-flash-preview (natančnost: 0.470; priklic: 0.650), pri čemer je posebej spodbudno, da smo za številne oblikoslovne probleme presegli F0.5 vrednost 0,7, ki nakazuje mejo zrelosti za aplikativno rabo. V pripravi je znanstveni prispevek, ki bo predstavil opisani napredek.

Naša zadnja naloga je bila priprava množic za evalvacijo strojnega popravljanja. Za evalvacijske množice smo zbrali besedila z avtentičnimi jezikovnimi napakami treh vrst: (A) besedila, ki so jih napisali mlajši govorci slovenščine, (B) besedila, ki so jih napisali odrasli govorci slovenščine in (C) besedila, ki so ji napisali odrasli govorci, ki se slovenščine učijo kot neprvega jezika. Glavni izziv zbiranja je bil, da besedila v popravljeni različici niso smela biti spletno dostopna, saj morajo evalvacijski podatki ostati razvijalcem skriti. Podatki bodo v nadaljevanju projekta označeni s petimi kategorijami: zapis, črkovanje, oblikoslovje, besedišče, skladnja. Za označevanje smo prilagodili specializirano označevalno orodje CJVT Svala. Ker morajo podatki ostati varovani, so množice v času priprave dostopne na zahtevo, do konca projekta pa jih bomo umestili tudi na platformo SloBench.

Kazalniki 2.2:

- [Veliki jezikovni modeli za preverjanje slovnice](#) (M18, na voljo).
- [Evalvacijske množice za preverjanje avtentične slovnice](#) (M24, dostopno pri Špela Arhar Holdt)

1.2.3. Aktivnost 2.3: Napredna slovnična analiza večjezičnih korpusov (M1-M36)

V aktivnosti T2.3 smo v prejšnjem obdobju izdelali veliki jezikovni model z izboljšanim slovničnim znanjem (M12), v tem obdobju pa smo si kot cilj zastavili podatkovno množico za evalvacijo slovničnega znanja velikih jezikovnih modelov (M18). To bo v naslednjem letu vodilo do več večjezičnih in medjezičnih slovničnih analiz (M36).

V prvem koraku smo sistematično evalvirali slovnično oziroma metajezikovno znanje velikih jezikovnih modelov. V ta namen smo razvili evalvacijsko množico [WALS-Bench](#) (Kazalnik 2.3.2-1), ki podatke iz zbirke *World Atlas of Language Structures* pretvarja v vprašanja o 192 jezikovnih značilnostih za več kot 2.600 jezikov. Evalvacija več sodobnih VJM je pokazala, da je njihovo eksplicitno znanje o jezikovni strukturi še precej omejeno ter izrazito neenakomerno med jeziki in jezikovnimi področji, zlasti pri manj zastopanih jezikih. Prispevek [Evaluating Metalinguistic Knowledge in Large Language Models across the World's Languages](#) z rezultati je sprejet v objavo v reviji *Natural Language Processing*.

Na tej osnovi smo razvili prvo različico agentnega sistema [UD-Agent](#) (Kazalnik 2.3.1), ki na slovnična vprašanja odgovarja s pomočjo dejanskih korpusnih podatkov iz večjezične zbirke *Universal Dependencies*. Sistem analizira uporabniško vprašanje, zgenerira in izvede ustrezno programsko kodo nad drevesnicami UD ter rezultate pretvori v odgovor v naravnem jeziku. Prvo različico smo evalvirali na množici [GramQA](#) v1 (Kazalnik 2.3.2-2), ki obsega 13 vprašanj o besednem redu za 179 jezikov. Primerjava z neposrednimi odgovori VJM brez dostopa do korpusnih podatkov je pri večini vprašanj

pokazala izboljšanje zanesljivosti; rezultati so predstavljeni v prispevku [Towards Corpus-Grounded Agentic LLMs for Multilingual Grammatical Analysis](#) (Klemen idr. 2026).

Trenutno sistem nadgrajujemo iz rešitve za omejen nabor vnaprej definiranih vprašanj v splošnejšega pomočnika za korpusno podprto slovnično analizo. Nova različica bo lahko sprejela čim širši nabor slovničnih vprašanj ter pripravila strukturiran odgovor s kratkim povzetkom, razlago, avtentičnimi korpusnimi primeri, grafičnimi ponazoritvami, opisom analitičnega postopka in priporočili za nadaljnje branje. Za evalvacijo razširjenega sistema smo razvili tudi novo različico evalvacijske množice GramQA 2.0, ki vključuje večje in vsebinsko precej raznovrstnejše število vprašanj s področij besedišča, morfologije, besednozvezne, stavčne in povedne skladnje.

Poleg zgoraj navedenih aktivnosti, povezanih z realizacijo projektnih kazalnikov, smo v okviru projekta nadgrajevali tudi raziskovalno infrastrukturo za slovnično analizo slovenskega jezika, kot so korpusi (SSJ, SST, UNER-SL), označevalni modeli (Stanza, Trankit) ter orodja za korpusno poizvedovanje (CQP-Tree, STARK, Drevesnik). Pripravili smo tudi pregled orodij za delo s podatki UD, predstavljen v prispevku za konferenco LREC, ter rezultate diseminirali na konferenci Grammatica Slavica. Organizirali smo tudi mednarodno multikonferenco [SyntaxFest 2025](#).

Kazalniki 2.3:

- Podatkovna množica za evalvacijo slovničnega znanja velikih jezikovnih modelov: [GramQA](#), [WALS Bench](#) (M18, na voljo).

1.3. Izziv 3: Veliki jezikovni modeli za govorjeni jezik

Veliki jezikovni modeli so zelo odvisni od podatkov. Pri rabi jezika obstajajo precejšnje razlike med različnimi vrstami diskurza. Večina jezikovnih podatkov izhaja iz pisnih virov. Obsežni viri govorjenih podatkov, če so sploh dostopni za raziskave, so običajno v lasti različnih podjetij ali institucij in so le redko na voljo raziskovalni skupnosti. Vendar pa se ti podatki, tudi če so javno dostopni, bistveno razlikujejo od govora, ki se uporablja v vsakdanjih pogovorih. Značilnosti, kot so ohlapnejše stavčne strukture, prekinjanje, motnje, molk, popravki, ponavljanja, napačna izgovorjava, pojasnjevanje, pritrdjevanje, konflikti, žaljivke in žargon (Yeomans idr., 2023), ponazarjajo površinske posebnosti pogovora in kažejo, da ima raziskovanje takih podatkov velik potencial vpliva na jezikovne raziskave (Love idr. 2014) ter predstavlja pomemben izziv za umetno inteligenco (Wahlster 2023). Pri tem izzivu želimo z najnovejšimi metodologijami zbiranja, filtriranja, avtomatskega transkribiranja in pragmatične obdelave govornih podatkov s pomočjo velikih jezikovnih modelov doseči napredek v govornih tehnologijah in z njimi povezanih raziskavah.

1.3.1. Aktivnost 3.1: Učinkovito zbiranje podatkov o govorjenem jeziku (M1-M36)

V prejšnjem letu izvajanja projekta smo zasnovali spletni vmesnik za zbiranje podatkov o pogovornem govoru (M4) ter v njem zbrali in ročno transkribirali pogovorne podatke za govorni učni korpus (5 ur) (M7). Cilj v poročanem obdobju je bil nadaljevanje zbiranja podatkov in njihova priprava za evalvacijo na portalu SloBench ASR v obsegu 1 ure ter dodani k obstoječim 3 uram (M25).

V poročanem obdobju je bil portal Govorjena slovenščina vsebinsko in funkcionalno nadgrajen ter uveljavljen kot infrastruktura za strukturirano zbiranje posnetkov vsakdanjega slovenskega govora z vključevanjem občanskih raziskovalcev. Koncept in metodologija portala sta bila predstavljena mednarodni raziskovalni skupnosti in objavljena v zborniku in v izbranih prispevkih letne konference CLARIN 2026 na Dunaju (DOI: 10.3384/ecp222.1597).

Konec septembra 2026 je zbranih približno 50 ur novih posnetkov zasebnih pogovorov iz vseh slovenskih statističnih regij. Objavljen je bil korpus ROG-dialog, portal pa je bil uporabljen tudi pri zbiranju podatkov za korpus Krohot in nadgradnjo zbirke SloBench ASR 2. Okrepljene so bile tudi aktivnosti za vključevanje javnosti: pripravljene so bile nove vsebine, navodila, novice, e-novičniki, videoposnetki in promocijske aktivnosti.

V teoretskem delu je bila pripravljena raziskava dejavnikov, ki vplivajo na pripravljenost za darovanje govornih posnetkov. Teoretični model povezuje teorijo samodoločanja (Self-Determination Theory, SDT), teorijo načrtovanega vedenja (Theory of Planned Behavior, TPB) in zaznano tveganje za zasebnost. Pripravljen je vprašalnik s 24 postavkami in sedmimi konstrukti, ki je metodološko pregledan in oddan v etično presojo. Opravljena je bila tudi sistematična raziskava motivacijskih dejavnikov v digitalnih in hibridnih projektih občanske znanosti, ki je zajela 16 študij in oblikovala priporočila za načrtovanje vključevanja prostovoljcev v zbiranje govornih podatkov; prispevek *Citizen science for spoken language resources: a systematic review of motivational factors influencing volunteer participation* je bil sprejet v objavo v reviji *Linguistics Vanguard* (<https://www.degruyterbrill.com/journal/key/lingvan/html>).

Kazalniki 3.1:

- Do konca drugega leta trajanja projekta (M24) ni bilo planiranega kazalnika.

1.3.2. Aktivnost 3.2: Semantično in pragmatično procesiranje govora (M7-M30)

Pri tej nalogi smo v obdobju od M12 do M24 zaključili označevanje korpusa ROG-dialog z dialoškimi dejanji in sentimentom, ki je bilo začeto v prvem poročevalskem obdobju. Opravljeno je bilo končno preverjanje in usklajevanje oznak, korpus pa je bil pripravljen za javno objavo. Korpus ROG-dialog 1.0, ki vsebuje posnetke, transkripcije, metapodatke ter ročne oznake dialoških dejanj in sentimenta, je bil objavljen v repozitoriju CLARIN.SI in je prosto dostopen na naslovu <http://hdl.handle.net/11356/2073>.

Pomemben del aktivnosti je bil namenjen označevanju podatkovnih množic SloBENCH1 in SloBENCH2 s sentimentom. Označevanje množice SloBENCH2 je bilo izvedeno z dvojnimi neodvisnimi označevanjem. Pri tem je bilo precej dela namenjenega zagotavljanju ustrezno visoke medoznačevalske skladnosti. Tako pridobljene anotacije predstavljajo kakovostne referenčne podatke za razvoj in vrednotenje avtomatskih metod za analizo sentimenta. Poleg tega smo zagotovili označenost podatkovne množice SloBENCH1 z izbrisljivimi elementi.

Na različnih pripravljenih in ročno označenih podatkovnih množicah smo izvedli prve preliminarne eksperimente avtomatskega napovedovanja sentimenta in dialoških dejanj. Namen eksperimentov je bil preveriti uporabnost pripravljenih anotacij za strojno učenje ter pridobiti začetne rezultate o zahtevnosti posameznih nalog in razlikah med podatkovnimi množicami. Rezultati teh eksperimentov predstavljajo osnovo za nadaljnji razvoj in sistematično vrednotenje modelov v naslednjem obdobju.

Rezultati dela so bili predstavljeni v več znanstvenih objavah. V članku Darinke Verdonik in Jasne Vidinič »Dialoška dejanja v zasebni govorni interakciji«, objavljenem v reviji Slavistična revija, so predstavljeni rezultati analize dialoških dejanj v zasebni govorni interakciji na podatkih korpusa ROG-dialog. Z raziskavami pojavov v govorjeni interakciji je povezano tudi delo Darinke Verdonik, Petra Rupnika in Nikole Ljubešića »Disfluencies in Public and Private Speech«, objavljeno v reviji Language and Speech, ki obravnava disfluentnosti ter njihove razlike med javnim in zasebnim govorom.

Kazalniki 3.2:

- Do konca drugega leta trajanja projekta (M24) ni bilo planiranega kazalnika.

1.3.3. Aktivnost 3.3: Nadzor kakovosti in filtriranje govornih podatkov (M1-M36)

Cilj naloge je izdelava zbirke slovenskih govornih posnetkov iz javno dostopnih virov namenjena avtomatskemu razpoznavanju govora (min. 300 ur) (M36).

V obdobju od M13 do M24 smo analizirali kategorije govora, ki vplivajo na uspešnost avtomatskega razpoznavanja govora (ASR), ter proučili vpliv govorcev, lastnosti govora in podatkovnih virov. Analizirali smo tudi akustične značilke za razvrstitev govorcev glede na različne kategorije, kot sta starost in spol, ter uporabili klasifikacijske postopke, ki temeljijo na metodah strojnega in globokega učenja.

Dokončali smo razvoj izhodiščnega slovenskega avtomatskega razpoznavalnika govora, ki bo služil za izhodišča pri zbiranju novih gradiv. Proučili smo vpliv različnih govornih slogov in čustvenih stanj govorcev na uspešnost avtomatskega razpoznavanja govora. Slovenski sistem ASR, temelječ na zbirki Artur, smo ovrednotili z uporabo slovenske zbirke emocionalnega govora Interface.

Analizirali smo robustnost sodobnih sistemov ASR pri domenskem in časovnem neujemanju govornih podatkov. Pri sistemih CrisperWhisper in WeNet smo izvedli analizo napak in parov zamenjav ter proučili značilnosti njunega razpoznavanja govora.

Izhodiščni avtomatski razpoznavalnik slovenskega govora smo uporabili kot privzeti razpoznavalnik za korpus Zila s 100 urami ročno transkribiranega govora slovenskega zilijskega narečja. Korpus Zila in iz izhodiščnega ASR sistema izpeljani narečni modeli so objavljeni v repozitoriju CLARIN.SI. Neobdelane zvočne posnetke smo preliminarно ovrednotili z uporabo metrik kakovosti govora.

Rezultate raziskovalnega dela smo v letu 2026 predstavili na več znanstvenih konferencah in v publikacijah (LREC 2026, SPECOM 2026, JTDH 2026, IEEE ELEKTRO 2026). Z izvedenimi aktivnostmi smo vzpostavili podlago za nadaljnjo pridobitev govornih podatkov iz odprtih virov, oceno njihove kakovosti, izdelavo obsežne slovenske govorne podatkovne zbirke ter njeno ovrednotenje z osnovnim sistemom ASR.

Kazalniki 3.3:

- Do konca drugega leta trajanja projekta (M24) ni bilo planiranega kazalnika.

1.3.4. Aktivnost 3.4: Veliki jezikovni modeli za prepoznavo domensko specifičnega govora (M1-M36)

Izvedli smo dodatne eksperimente o plitvi integraciji modelov avtomatske razpoznavne govora in velikih jezikovnih modelov (VLM) ter obsežno študijo pristopov implementacije globljih metod integracije govora v velike jezikovne modele SALM (Speech-augmented Language Model). Pri plitvih metodah smo evalvirali metode ponovnega ocenjevanja hipotez, popravljanje napak z generativnimi modeli, izborom najboljše hipoteze ter prilagajanja generativnih modelov za transformacijo hipotez razpoznavalnika v pravičen zapis. Pri globoki integraciji smo se osredotočili predvsem na metodo SALM, kjer VJM deluje kot dekode in je potrebno učiti projekcijske sloje, ki transformirajo avdio vložitve razpoznavalnika v tekstovne vložitve VJM. Pri tem smo izvedli tudi eksperimente s številom in obliko projekcijskih slojev ter uporabo metod prilagajanja, kot je LoRA.

Evalvirali smo devet VJM (SloBERTa, GaMS-9b, EuroLLM, Gemma 2 9b, Gemma 2 27b, Gemma 3 9b, Gemma 3 27b, Llama 3.1 8b in SlovenianGPT), na šestih množicah (Artur 1.0, Gos VideoLectures 4.2,

CommonVoice 18.0 SL, VoxPopuli SL, FLEURS SL in Sofes 1.0). Raznolike evalvacijske množice nudijo vpogled v delovanje pristopov in modelov na raznolikih govorcih, stilih govorjenja in označevanja transkriptov. Tekom eksperimentov smo spremljali tudi kako količina videnih hipotez vpliva na uspešnost metod.

Rezultati kažejo, da metoda SALM trenutno ne izboljša rezultatov metrike WER na različnih podatkovnih množicah. Kot pri ostalih generativnih pristopih (npr. pri popravljanje jezikovnih napak), VJMji pogosto dodajo ali spremenijo stavčno strukturo transkripta, kar je za trenutno nalogo (dobeseden, pravilen transkript) narobe.

Trenutno najboljša metoda ponovnega ocenjevanja hipotez tako ostaja model SlovenianGPT, ki pa mu tesno sledi metoda prilagajanja generativnih modelov za transformacijo hipotez razpoznavalnika v pravilen transkript.

Kazalniki 3.4:

- Do konca drugega leta trajanja projekta (M24) ni bilo planiranega kazalnika.

1.4. Izziv 4: Napredne tehnologije za digitalno humanistiko

Specifične pomembne naloge na področju digitalne humanistike zahtevajo nove napredne jezikovne tehnologije ali prilagoditve obstoječih splošnih pristopov. V tem delu poročila predstavljamo štiri sklope tehnologij za reševanje problemov, ki služijo kot osnova specifičnim nalogam digitalne humanistike, ki jih obravnavamo v izzivu 5.

1.4.1. Aktivnost 4.1: Grafi imenskih entitet (M1-M18)

Pripravili smo zbirko ročno označenih imenskih entitet v zgodovinskih parlamentarnih dokumentih zbirke Kranjska 1.0. Objavljena bo v okviru zbirke Kranjska 2.0, ki bo prav tako vsebovala samodejno označene imenske entitete s pomočjo jezikovnega modela Gemma 4.

Na podlagi označenih podatkov smo v prejšnjem letu razvili interaktivno vizualizacijo grafov imenskih entitet (D4.1, M12), ki prikazuje omrežja ljudi, organizacij, krajev in drugih entitet ter njihove medsebojne povezave. To vizualizacijo smo, prek povratnih informacij s strani domenskih raziskovalcev, razširili z vizualizacijo negotovosti označb, tako na nivoju entitete in njenih atributov kot povezav. Razširjen sistem omogoča večkratne povezave kot simulacijo entitet z več interpretacijami, urejanje podatkov, dodajanja novih vnosov in naprednejše funkcije filtriranja.

Kot predstavitev novega sistema za vizualizacijo pripravljamo članek z naslovom *EntityWeaver: Visual Exploration and Curation of Named-Entity Relationships in Document Collections*, ki potrebuje še poglobljeno evalvacijo.

Izvedli smo poglobljeno evalvacijo učinkovitosti različnih velikih jezikovnih modelov za prepoznavanje imenskih entitet, sentimenta na nivoju kolektivnih identitet na izbranih zgodovinskih korpusih in zaznavanja dogodkov v slovenskih zbirkah. Pokazalo se je, da se modeli na zgodovinskih podatkih obnesejo precej slabše kot na novejših, še posebej pri zaznavanju entitet, ki označujejo organizacije. Članek o imenskih entitetah smo objavili na konferenci JT-DH 2026, članek o sentimentu pa na delavnici LLMS4SSH 2026, ki se je izvajala v okviru konference LREC 2026. Članek o zaznavanju dogodkov nameravamo objaviti v kratkem.

Kazalniki 4.1:

- [Visualization of extracted named entity graphs](#) (M24, na voljo).

1.4.2. Aktivnost 4.2: Veliki jezikovni modeli za diahrono analizo (M6-M24)

Cilj naloge je razviti nove pristope za avtomatizirano odkrivanje semantičnih sprememb z uporabo velikih jezikovnih modelov, ki bodo omogočili diahrono analizo tudi pri subtilnih pomenskih razlikah.

Pristop, ki smo ga razvili pred začetkom projekta (Pranjić et al., 2024) in temelji na vložitvah modela SloBERTa, smo nadgradili s teoretično analizo vpliva razpršenosti vložitev na zaznavanje pomenskih sprememb. Prispevek je bil sprejet za objavo v reviji PeerJ.

Na podatkovni množici, ki predstavlja slovenski zlati standard smo preverili, kako enajst generativnih velikih jezikovnih modelov (VJM) ocenjuje pomensko podobnost brez dodatnega učenja. Primerjali smo jih z človeškimi ocenami ter z manjšimi kodirnimi modeli. Medtem ko lahko s pristopi z VJM pridobimo uporaben signal o pomenski podobnosti rab, prepričljive prednosti pristopa z VJM pred manjšimi modeli tipa BERT nismo zaznali.

Nadaljnje delo smo zato usmerili v modele s kodirnikom (tipa BERT) ter namensko učene vektorske vložitve, ki so še posebej primerne za analizo velikih korpusov digitalne humanistike. Poleg modela, ki temelji na vložitvah SentenceBERT, smo razvili pristop z namensko prilagojenim modelom za predstavitev rab, sorodnim modelu XL-LEXEME. Oba pristopa uporabljata gručenje semantičnih rab na podlagi grafov. Pristopi so zasnovani tako, da lahko zgodovinarji preučujejo različna časovna obdobja, posamezne časopise ali druge izbrane dele gradiva.

Pristope smo preizkusili na dveh področjih: na večjezični zbirki novic iz let 2005–2024, v kateri smo spremljali rabo izraza »globalno segrevanje«, ter na zgodovinskem korpusu slovenskih periodičnih publikacij (sPeriodika 1.0), kjer smo spremljali rabo izraza »revščina«.

Kazalnik 4.2:

- [Novel methodology for diachronic analysis using LLMs](#) (M24, na voljo).

1.4.3. Aktivnost 4.3: Multimodalni modeli za analizo slik (M10-M30)

V okviru naloge, osredotočene na prilagoditev slikovno-jezikovnih modelov (SJM) za uporabo v digitalni humanistiki, sta zastavljena dva glavna cilja: priprava podatkovne množice slik iz slovenskega časopisa iz starejših zgodovinskih obdobj (M24) in razvoj slikovno-jezikovnega modela, prilagojenega za izbrane aktivnosti digitalne humanistike (M30).

Prvi cilj smo uspešno izpolnili z vzpostavitvijo prve multimodalne (slikovno-tekstovne) primerjalne podatkovne zbirke za zgodovinsko slovensko časopisje, ki je dostopna na platformi [CLARIN](#). Zbirka je bila zgrajena avtomatsko s pomočjo modelov SAM3 za segmentacijo regij in Gemini 3 Pro Preview za multimodalno označevanje, ter preverjena s pomočjo ročne evalvacije. Zbirka omogoča učenje in evalvacijo SJM pri nalogah luščenja besedila in prostorske lokalizacije vizualnih elementov v zgodovinskih časopisih. Evalvacija šestih SJM je razkrila vrzel med uspešnim prepoznavanjem besedila in znatno zahtevnejšim prostorskim umeščanjem, hkrati pa je potrdila, da nadzorovano prilagajanje

(angl. *supervised fine-tuning*) modelov bistveno izboljša strukturirano razumevanje strani in prostorsko umeščanje vizualnih elementov.

Drugi pomemben dosežek predstavlja znanstveni članek, objavljen v mednarodni reviji *International Journal on Document Analysis and Recognition (IJ DAR)* z naslovom »*Large language models for OCR in cultural heritage: a comparative study on Slovene Folkloristic texts*«. Raziskava prinaša prvo sistematično primerjalno študijo sodobnih pristopov optične prepoznave znakov (OCR) na korpusu slovenske kulturne dediščine, pri čemer je bilo na tipkopisnih folklornih rokopisih ter vizualno heterogenih izdajah revije Ciciban primerjanih šest različnih metod. Rezultati kažejo, da neposredni multimodalni sistemi dosegajo najvišjo natančnost na tipkopisih, medtem ko je uspešnost pri zahtevnejših postavitvah močno odvisna od strukture strani, kar poudarja nujnost vsebinsko prilagojenih postopkov za potrebe digitalne humanistike.

Kazalniki 4.3:

- [Podatkovna množica slik iz slovenskega časopisja iz starejših zgodovinskih obdobj](#) (M24, na voljo)

1.4.4. Aktivnost 4.4: Veliki jezikovni modeli in RAG za detekcijo kontradikcij (M1-M18)

Protislovja v zakonodaji in sodni praksi predstavljajo resen problem za pravno varnost, konsistentnost in interpretacijo prava. Vzpostavitev sistema, ki bo avtomatsko prepoznaval vsebinska neskladja, zato predstavlja pomemben korak k učinkovitejši pravni analizi in razvoju naprednih pravnih informacijskih sistemov in kvalitetnejši pravni nomotehniki.

Za potrebe razvoja domensko prilagojenih vložitenih modelov smo izvedli večjo označevalno kampanjo za pripravo množice sklepanja v naravnem jeziku (angl. *Natural Language Inference, NLI*). Množico so pripravili štirje označevalci, magistrski študenti prava. Množica je sestavljena iz 2.214 parov in pokriva štiri zakone: Stvarnopravni zakonik (SPZ), Oblgacijski zakonik (OZ), Kazenski zakonik (KZ) in Zakon o varstvu osebnih podatkov 2 (ZVOP-2). Vsi pari so bili ročno pregledani s strani domenskih ekspertov. Množica je prosto dostopna in javno objavljena na repozitoriju Clarin.

Korpus COLESLAW, pripravljen v sklopu naloge T5.3, in SLawNLI smo uporabili za razvoj sistema za zaznavo pravnih neskladij. Razviti sistem ContrAG temelji na domensko prilagojenem vložitenem modelu ter domensko prilagojeni različici modela GAMS3-12B-Instruct in je prilagojen za iskanje pravnih neskladij na zakonodajnem nivoju z metodo pridobivanja (angl. *retrieval*) in prerazporejanja (angl. *reranking*).

Na domenski zakonodaji, prisotni v množici SLawNLI, sistem ContrAG dosega rezultat po metriki acc@10 0.976, kar pomeni, da zanesljivo odkrije neskladne člene. Ko preiskovalni prostor razširimo na celotno aktivno zakonodajo (več kot 7000 zakonov), metrika acc@10 pade na 0.637. Degradacija je pričakovana, saj je dokumentov več kot 1700x več (4 zakoni nasproti celotni zakonodaji).

Za predstavitev dosedajnega dela na sistemu ContrAG pripravljamo prispevek za revijo *Artificial Intelligence and Law*. V nadaljevanju bomo nadaljevali razvoj sistema ContrAG upoštevajoč pravno hierarhijo, razširitev množice SLawNLI 1.0 in korpusa COLESLAW.

Kazalniki 4.4

- [Novi RAG sistem za slovenščino, zmožen detekcije protislovij v dokumentih](#) (M18, na voljo).

1.5. Izziv 5: Izbrani izzivi digitalne humanstike

1.5.1. Aktivnost 5.1: Veliki jezikovni modeli za historiografijo (M1-M36)

V drugem letu projekta smo težišče dela prenesli z metodoloških priprav na izvedbo analiz in evalvacij. Rezultate smo predstavili v osmih prispevkih: dva na delavnicah konference LREC 2026, tri na konferenci Jezikovne tehnologije in digitalna humanistika (JT-DH 2026), en članek je v recenzijskem postopku v reviji *Prispevki za novejšo zgodovino*, en povzetek je v presoji za konferenco [QDLCA27](#) in še eden pri konferenci *Impresso: Beyond Borders - Futures of Digitised Historical Media Collections for Research and Teaching*.

Na področju novih metodoloških pristopov k zgodovinski analizi ideoloških konceptov smo predstavili dve študiji (obe na podlagi korpusa sPeriodika). V prispevku [Thematic Landscapes of the Past](#), predstavljenem na delavnici PressMint, smo s tematskim modeliranjem orisali tematsko sestavo slovenskega zgodovinskega periodičnega tiska. V prispevku [Benchmarking LLMs for Aspect-Based Sentiment Classification in Slovene Historical Periodicals](#), predstavljenem na delavnici LLMs4SSH, pa smo evalvirali šest odprtih velikih jezikovnih modelov (VJM) pri prepoznavanju odnosa do omemb kolektivnih identitet v besedilih, pridobljenih z OCR. Najboljše rezultate je dosegel slovenščini prilagojeni model GaMS3-12B-Instruct (F1 0,708). Analiza je pokazala sistematične razlike v uspešnosti: pozitivni odnos ostaja podznan, zlasti pri samostalniških in skupinskih omembah, negativni pa je prekomerno pripisan. Ročno označeno zbirko 371 omemb smo objavili v [repozitoriju projekta](#), napovedi najboljšega modela pa smo nato razširili na 2,65 milijona omemb in jih agregirali po identitetah in časopisih. Izsledke smo nadgradili v prispevku [Evaluative Framing of Collective Identities in Leading Slovene Newspapers from the Late Habsburg Empire](#) na konferenci JT-DH 2026.

Delo na grafih imenskih entitet smo predstavili v prispevku [Combining Graph Visualisation and Critical Discourse Analysis in the Study of Slovene Historical Newspapers](#) na konferenci JT-DH 2026, obsežnejšo različico analize pa smo oddali reviji *Prispevki za novejšo zgodovino*.

Na področju diahronih študij konceptov smo po razvoju testirali in ocenjevali spletno orodje [Semantic Shift Viewer](#), ki omogoča zaznavanje in vizualizacijo pomenskih premikov v korpusih sPeriodika, Kranjska in yu1Parl ter združuje pristope na ravni posameznih pojavitev, na ravni skupin pomenov in na ravni dodatno naučenih modelov. Orodje in primerjavo šestih vektorskih modelov na ključnih besedah kriza, korupcija in jezik smo predstavili v prispevku [Tracking Lexical Semantic Change in Historical Slovenian Corpora with Contextual Embeddings](#) na konferenci JT-DH 2026. Izkazalo se je, da noben model ni dosledno najboljši v vseh merilih, zato priporočamo primerjavo več modelov hkrati. Na tej podlagi smo za konferenco QDLCA27 pripravili povzetek, v katerem preverjamo, ali so različni tipi pomenske spremembe (razširitev, zožitev, izboljšanje, poslabšanje, metafora in metonimija) enako zaznavni z obstoječimi metodami.

Kazalniki 5.1:

- [Novi metodološki pristopi k zgodovinski in ideološki analizi z uporabo velikih jezikovnih modelov](#) LLMs (M18, na voljo).

1.5.2. Aktivnost 5.2: Veliki jezikovni modeli za folkloristiko (M1-M36)

Cilj naloge je razviti metodologijo, ki temelji na velikih jezikovnih modelih (VJM), za računsko folkloristiko. Namen je povečati učinkovitost in natančnost primerjalnih raziskav v folkloristiki ter omogočiti obsežne diahrone, medjezikovne in medkulturne primerjave. Pri tem izhajamo iz dejstva, da

pravlјice, ljudske pripovedi in druge etnografske naracije niso zgolj brezčasne zgodbe, temveč odsevi dru֓b, ki so jih ustvarjale in posredovale. Osrednji izziv je zato avtomatizacija analize pripovednih struktur ob preuĉevanju razliĉnih razliĉic zgodb skozi ĉas in kulture.

V poroĉenem obdobju smo nadgradili delo na űtirih kljuĉnih digitalnih korpusih: (1) korpusu etnografskih besedil, (2) korpusu pravlјic iz zbirke ZRC SAZU, (3) korpusu revije Ciciban in (4) prilagojenem korpusu Kmetijskih in rokodelskih novic (KRN). Nadaljevali smo razvoj in uporabo postopkov za transkripcijo, normalizacijo, strukturiranje in raĉunalniűko analizo folkloristiĉnega gradiva. V izroĉku D5.2.1 smo razvili novo metodologijo za digitalno folkloristiko, strukturirano v treh fazah: vzpostavljanje zanesljivih podatkovnih baz, anotacija podatkov in metapodatkov ter napredna analitiĉna obdelava. Nadaljnje aktivnosti so bile usmerjene v uporabo, evalvacijo in nadgradnjo razvitih postopkov na pripravljenih korpusih.

Pomemben del aktivnosti je bil namenjen pripravi digitaliziranega korpusa revije Ciciban. Izvedena je bila avtomatizirana segmentacija celotnega korpusa, ki obsega 46 letnikov, pribliűno 10.000 strani in 20.505 avtomatizirano identificiranih segmentov. Roĉna evalvacija segmentacije in klasifikacije na vzorcu 52 enot je pokazala teűave predvsem pri doloĉanju meja posameznih besedil ter klasifikaciji tipov besedil. Nadaljnje delo je zato usmerjeno v izboljšanje in evalvacijo segmentacije ter razvoj klasifikacije besedil. Razvit in preizkuűen je bil tudi cevovod SAM3–Gemini za ekstrakcijo slik, pripravo opisov slik, doloĉanje omejitvenih okvirjev (bounding boxes) in segmentacijo strani.

Raziskave so potekale tudi na podroĉju avtomatizirane analize folkloristiĉnih motivov in pripovednih struktur. Avtomatizirano prepoznavanje motivov in tipov pravlјic na primeru Pepelke je bilo nadgrajeno z raziskavami prepoznavanja razliĉnih tipov junakov v ljudskih pripovedih, vkljuĉno z evalvacijo velikih jezikovnih modelov pri klasifikaciji lika izobĉenca (angl. outlaw hero). Nadaljevali smo delo na podroĉju strukturno-semiotiĉne in multimodalne analize, zlasti pri uporabi semiotiĉnega kvadrata za analizo pripovedi o zmajih ter raziskovanju kulturno-semiotiĉnega sklepanja slovenskih vizualno-jezikovnih modelov.

Na podroĉju űivalskih pripovedi so bile izvedene raziskave moralnega diskurza in kulturnega spomina v slovenskih űivalskih pripovedih v űirűem slovanskem pripovednem izroĉilu. Po pripravi in segmentaciji korpusa Ciciban poteka tudi raziskava politizacije űivali v reviji Ciciban.

Na podroĉju praks reűevanja konfliktov smo pripravili veĉjeziĉni korpus 108 zgodovinskih, protoetnografskih in antropoloűkih del, ki obsega pribliűno 16,9 milijona besed. Besedila smo oĉistili in razdelili na 23.905 analitiĉnih enot. Korpus smo sistematiĉno analizirali z velikim jezikovnim modelom na podlagi veĉkrat preizkuűenega in roĉno preverjenega raziskovalnega poziva za prepoznavanje praks in ritualov reűevanja konfliktov. Rezultate modela smo uporabljali kot orodje za iskanje in razvrűĉanje relevantnih odlomkov, pomembnejűe ugotovitve pa smo preverili z neposrednim branjem virov. Na tej smo pripravili ĉlanek *Peace as Strong as Oak: Ritual Law, Vendetta, and Mediation in Renaissance Europe and the Americas*, ki je űe objavljen kot prednatis, ter drugi ĉlanek o najzgodnejűih poroĉilih o Amerikah in obiĉajnem reűevanju konfliktov, ki je v zakljuĉni fazi priprave. Raziskave so neposredno povezane z naslednjim predvidenim izroĉkom o ritualih reűevanja konfliktov (M36).

V poroĉanem obdobju je potekalo tudi mednarodno raziskovalno sodelovanje v okviru COST akcije CA22149 CHANGECODE, bilateralnega projekta ARIS BI-US/24-26-034 ter projekta AI4DH. Raziskovalna gostovanja in sodelovanje z mednarodnimi raziskovalnimi institucijami so bila usmerjena v razvoj in uporabo metodoloűkih pristopov za raĉunalniűko analizo folkloristiĉnega gradiva, medkulturno primerjalno raziskovanje pravlјic ter razvoj metod in infrastrukture za digitalno humanistiko.

Doseženi rezultati predstavljajo nadaljevanje in nadgradnjo dela iz prvega leta projekta ter osnovo za nadaljnjo uporabo, evalvacijo in nadgradnjo metodologije. Segmentacija in klasifikacija folklorističnega gradiva, razvoj multimodalnih postopkov ter raziskave motivov, pripovednih struktur in praks reševanja konfliktov so usmerjeni v nadaljnje analize in pripravo naslednjega predvidenega izročka (M36).

Kazalniki 5.2:

- [Veliki jezikovni modeli za folkloristiko](#) (M18, na voljo)

1.5.3. Aktivnost 5.3: Veliki jezikovni modeli in RAG za pravno domeno (M1-M36)

V sklopu naloge T5.3 smo pripravili največji korpus slovenskega pravnega jezika COLESLAW, pozetega iz spletnih virov pravnih besedil (PISRS, USRS, Sodna praksa in Državni list). Prečiščen korpus vsebuje okoli 540 tisoč dokumentov ter približno 770 milijonov besed ter služi kot vir znanja za nadaljnje aktivnosti. Korpus je prosto dostopen in javno objavljen na repozitoriju Clarin.

Kot primer uporabe korpusa smo z uporabo množice SLawNLI, razvite v aktivnosti T4.4, prilagodili vložitvene modele bge-m3 in embedding-gemma za splošno iskanje dokumentov. Te uporabljamo za razvoj splošnega pravnega klepetalnika, ki dostopa do aktivnih predpisov v korpusu COLESLAW.

Poleg pravnega klepetalnika in sistema za zaznavo pravnih neskladij smo identificirali še dva primera uporabe razvitih modelov in pripravljenih korpusov. V prvem na podlagi opisa podjetja in aktivnosti, ki jih opravlja, preiskujemo predpise, zakone in odredbe, ki so za to podjetje relevantna. Cilj aktivnosti je razviti sistem obveščanja ter spremljanja zakonodaje, ki določa delovanje podjetja. Opravljeni uvodni eksperimenti glede priprave vložitev zakonodaje so pozitivni, v sklopu 20 odkritih zakonov/predpisov/uredb jih je približno 65% pravih.

Drug primer uporabe je zaznava in razlaga pravne terminologije. Pravni jezik se razlikuje od splošne slovenščine in pogosto uporablja specifične ter težko razumljive izraze.

Seznam rešitev oz. primerov uporabe v nalogi T5.3 je torej: zaznava pravnih neskladij, pravni asistent oz. klepetalnik, identifikacija relevantne zakonodaje za podjetja ter zaznava in razlaga pravne terminologije.

O izvedenih aktivnostih smo pripravili dva prispevka v reviji Pravna Praksa.

Kazalniki 5.3:

- [Database for Slovene legal texts](#) (M18, na voljo).

1.6. Izziv 6: Evalvacija in razumevanje velikih jezikovnih modelov

Primerjalno testiranje je ključni instrument za merjenje in spremljanje uspešnosti velikih jezikovnih modelov, ki ves čas dobivajo nove različice (Zhao et al. 2023). Čeprav imajo statični primerjalni testi omejitve, med katerimi je najpomembnejša kontaminacija velikih jezikovnih modelov s podatki iz primerjalnih testov (Zhou et al. 2023), so še vedno najbolj izvedljiva rešitev za spremljanje uspešnosti za jezike z manjšim številom govorcev, za katere so dinamične alternative, kot so arene velikih jezikovnih modelov (Chiang et al. 2024), manj primerne. Za boljše razumevanje velikih jezikovnih modelov potrebujemo primerjalne teste, ki so boljši pri merjenju višjih ravni razumevanja, zlasti povezanih s kompleksnim izražanjem, figurativnim jezikom in govorjenim jezikom. Enako pomembno je odkrivanje pristranskosti v velikih jezikovnih modelih in neposredne razlage vedenja velikih jezikovnih modelov pri pomembnih nalogah.

1.6.1. Aktivnost 6.1: Figurativni jezik in pragmatični benčmarki (M1-M24)

Figurativni jezik, ki vključuje metafore, ironijo in sarkazem, je pomembna značilnost človeške komunikacije. Čeprav veliki jezikovni modeli kažejo izjemne sposobnosti pri prilagajanju določenemu slogu in ustvarjanju inovativnih metafor, je njihova razen razumevanja sarkazma in ironije še vedno nizka (Yakura 2024). Težave imajo tudi z razumevanjem ali detektiranjem posrednih prošelj in izrazov faux pas (Strachan et al. 2024). Prav tako veliki jezikovni modeli bistveno zaostajajo za ljudmi pri svojih pragmatičnih sposobnostih, kot so upoštevanje zunajjezikovnega konteksta, namenov govorca, predpostavk in implicitnih pomenov (Srvanthi et al. 2024). Glavna težava pri ocenjevanju zmožnosti velikih jezikovnih modelov pri niansiranem razumevanju in tvorjenju jezika je pomanjkanje evalvacijskih množic in primerjalnih testov, zato smo si zastavili več ciljev, s katerimi zapolnujemo vrzeli na področju evalvacijskih množic za slovenščino, in sicer za področje pragmatičnega pomena in figurativnega jezika.

V poročenem obdobju M13-M24 smo nadaljevali diseminacijo in eksperimente z množicami za razumevanje metafor, ironije in sarkazma SloPragMega ter pragmatičnega vedenja SloPragEval, ki smo ju izdelali v prvem letu projekta. Podatkovni množici in pilotne eksperimente vrednotenja jezikovnih modelov smo predstavili na delavnici RESOURCEFUL, ki je potekala ob mednarodni konferenci LREC26. Rezultati kažejo, da so s prevodi iz angleščine ustvarjene množice in vprašanja zaprtega tipa postala relativno enostavno rešljiva z novejšimi jezikovnimi modeli.

Poleg slovenske različice smo primerjalni test PragMega še dodatno razširili, in sicer smo ga prevedli in lokalizirali v češčino, slovaščino in hrvaščino. Trojezična različica PragMega CzeSloCro je javno dostopna v repozitoriju Clarin.si. <http://hdl.handle.net/11356/2261>.

V tem obdobju smo izdelali novo podatkovno množico z asociacijami in razlago asociativnega vedenja WAX-SL. Množica je bila izdelana po enaki metodologiji kot angleška množica asociacij in razlag WAX-EN, in sicer so poleg podajanja asociacij človeški označevalci dodali še razlage asociacij, nekatere od teh pa so še dodatno označene z relacijami med iztočnicami in asociacijami. Slovenski WAX-SL tako vsebuje 1.262 trojic (iztočnica, asociacija, razlaga), od katerih ima 700 še oznako relacije. Rezultat je dostopen v repozitoriju CLARIN.SI na <http://hdl.handle.net/11356/2358>.

Množica ter prvi eksperimenti, v katerih primerjamo človeško asociativno vedenje z asociativnim vedenjem jezikovnih modelov, bodo v oktobru 2026 predstavljeni s prispevkom na delavnici LUHME, ki poteka ob konferenci EMNLP 2026. Prvi rezultati kažejo, da se človeško asociativno vedenje kvalitativno razlikuje od asociacij, ki jih podajajo jezikovni modeli.

Kazalnik 6.1:

- [Primerjalni test razlage pragmatičnega in asociativnega vedenja](#) (M24, na voljo).

1.6.2. Aktivnost 6.2: Primerjalni test razumevanja govorjenega jezika (M25-M36)

Obstajajo številni primerjalni testi razumevanja pisnega jezika, primerjalnih testov razumevanja govorjenega jezika pa je precej manj. Z vse več multimodalnih velikih jezikovnih modelov, ki podpirajo govor (Barrault et al. 2023, Hu et al. 2024, Fathullah et al. 2024), se potreba po primerjalnih testih razumevanja govorjenega jezika znatno povečuje. Cilj aktivnosti je razviti primerjalni test razumevanja govorjenega jezika za slovenščino in ga vzpostaviti v evalvacijski platformi SloBENCH za evalvacijo uspešnosti sedanjih in prihodnjih velikih jezikovnih modelov, ki podpirajo govor.

Cilji naloge so izdelava govorne množice (4 ure) z anotacijami govornega dejanja in sentimenta, večreferenčne evalvacije za avtomatsko razpoznavo govora in procesiranja dialoga in sentimenta v govoru.

V obdobju od M13 do M24 smo nadaljevali razvoj testnih množic in metodologije za primerjalno vrednotenje sistemov za avtomatsko razpoznavanje govora. Na podlagi podatkov, zbranih in označenih v okviru projekta, smo pripravili novo testno množico SloBENCH2, namenjeno zahtevnejšemu vrednotenju razpoznavalnikov slovenskega spontanega govora. V primerjavi z obstoječo množico SloBENCH1 vsebuje novi SloBENCH2 zahtevnejši govor, posnet v naravnih in akustično manj nadzorovanih okoliščinah, z več narečnega in pogovornega govora. Podatki vsebujejo dva nivoja prepisa, pogovornega in normaliziranega, ter posebej označene izbrisljive elemente, predvsem govorne netekočnosti, katerih vključitev oziroma izpust ne spremeni bistveno vsebine izrečenega.

Razvili in preizkusili smo večreferenčni pristop k vrednotenju razpoznavalnikov govora, ki upošteva, da za spontani govor pogosto ne obstaja en sam pravilen prepis. Opredelili smo sedem načinov tvorjenja referenčnih prepisov, ki omogočajo izbiro med pogovorno in normalizirano obliko na ravni celotnega izreka ali posamezne besede ter ustrezno obravnavo izbrisljivih elementov. Pristope smo sistematično ovrednotili na štirih različnih sodobnih sistemih za razpoznavanje govora. Rezultati kažejo, da lahko izbira načina prepisa in obravnava govornih netekočnosti pomembno vplivata na standardni metriki WER in CER. Pri posameznih sistemih večreferenčno vrednotenje zmanjša izmerjeno napako do 20 % glede na klasično vrednotenje z eno normalizirano referenco. Pokazali smo tudi, da večreferenčno vrednotenje omogoča profiliranje sistemov glede na njihovo nagnjenost k normalizaciji pogovornega govora ter k ohranjanju oziroma izpuščanju govornih netekočnosti. Dodatno smo pokazali, da tako optimizirani referenčni prepisi dosegajo statistično značilno višjo korelacijo standardnih metrik WER in CER s semantičnimi ocenami na podlagi velikih jezikovnih modelov kot običajno enoreferenčno vrednotenje.

Kazalnik 6.2:

- V drugem letu trajanja projekta (M24) kazalnika ni bilo predvidenega.

1.6.3. Aktivnost 6.3: Detekcija pristranskosti v velikih jezikovnih modelih in pri avtomatski prepoznavi govora (M10-M36)

Množico za analizo pristranosti na podlagi spola in etnične pripadnosti (EEC-SL) smo uporabili za analizo pristranosti velikih jezikovnih modelov, kjer so manjši modeli pokazali majhno etnično pristranost. Prispevek je bil objavljen na konferenci LREC (Caporusso et al., 2026).

Lokalizirali smo množico za evalvacijo moralnega presojanja jezikovnih modelov (Novak, 2026) ter primerjali etično presojanje med ljudmi in jezikovnimi modeli (Machidon et al. 2026). Analizirali smo politično pristranost modelov in članek o tem objavili na konferenci LTDH (Debevc & Martinc, 2026).

Analizirali smo družbene pristranskosti (npr. glede spola in rase) v večmodalnih modelih in pokazali, da ima lahko vir učnih podatkov večji vpliv na pristranost kot velikost modela (Al Sahili et al., 2025a). Pri večjezičnih modelih so pristranskosti pogosto izrazitejše v jezikih z manj viri, pristranskosti iz jezikov s slovničnim spolom pa se lahko prenesejo tudi v jezike brez slovničnega spola; članek o tem, objavljen na konferenci AACL (Al Sahili et al., 2025b), je prejel priznanje "častna omemba".

Razvili smo tudi pristop za razpristranjevanje multimodalnih modelov na nalogi generiranja slik iz besedil (Al Sahili et al., 2025c).

Kazalniki 6.3:

- [Detekcija pristranskosti za slovenščino](#) (M24).

1.6.4. Aktivnost 6.4: Razlage velikih jezikovnih modelov na osnovi znanja (M19-M36)

Transparentnost in zanesljivost velikih jezikovnih modelov sta temeljnega pomena za krepitev zaupanja, zagotavljanje pravičnosti in etične uporabe, spodbujanje inovacij ter skladnost z regulacijskimi standardi. Čeprav obstajajo delne rešitve pojasnjevanja, ki večinoma temeljijo na samopojasnjevanju, to ni dovolj. Za naslovitev problema pomanjkanja transparentnosti velikih jezikovnih modelov smo razvili splošno metodologijo, ki smo jo testno uporabili na področju zdravorazumskega sklepanja.

Zdravorazumsko sklepanje je pomemben in zahteven vidik razumevanja naravnega jezika. Winogradova shema (WSC) je široko uporabljena podatkovna množica, zasnovana za testiranje takšnega sklepanja na podlagi pomena zaimkov. Čeprav lahko številni veliki jezikovni modeli dajo pravilne odgovore, so veliko šibkejši pri razlagah, odgovorov. Izvedli smo začetno študijo na množici KE-WSC, ki je razširjena s človeškimi razlagami. Pri analizi smo ocenjevali vpliv vključevanja človeških razlag na delovanje VJM. S pomočjo eksperimentov na različnih družinah in velikostih modelov smo ugotavljali, ali dodatne razlage, bodisi podane ali dodatno generirane, izboljšujejo zdravorazumsko sklepanje. Naši eksperimenti kažejo, da modeli brez sposobnosti sklepanja pri tej nalogi slabo delujejo., medtem ko modeli s sklepanjem na podlagi podane razlage presegajo osnovne modele. Delo smo poslali v recenzijo v revijo.

Kazalnik 6.4:

- Kazalnik »Nova metodologija za razlage na osnovi znanja za velike jezikovne modele« je predviden na koncu projekta (M36).

1.7. Delovni sklop 7: Upravljanje in diseminacija

Delovni sklop je namenjen usklajevanju in spremljanju napredka projekta, vključno z upravljanjem etičnih vprašanj, tveganj in zagotavljanjem kakovosti. Poseben poudarek je na ravnanju z raziskovalnimi podatki v skladu z načeli odprte znanosti. Poleg tega ta projektni sklop skrbi za dejavnosti komuniciranja, diseminacije in eksploatacije, s ciljem povečati prepoznavnost projekta med različnimi deležniki.

1.7.1. Aktivnost 7.1: Koordinacija in spremljanje projekta (M1-M36)

Koordinacija projektnih aktivnosti je potekala v okviru rednih sestankov članov delovnih aktivnosti in rednih mesečnih sestankov upravnega odbora za spremljanje napredkov projekta. Izdelali smo dokument z aktivnostmi obvladovanja tveganj, s katerim so se seznanili vsi raziskovalci. Polletno izvajamo srečanja vseh sodelavcev projekta, kjer predstavimo dosežene rezultate, usklajujemo delo med raziskovalnimi sklopi in skupinami ter pripravimo načrt dela za naslednje polletno obdobje. Do sedaj so bila izvedena štiri tovrstna srečanja (4. 11. 2024 na INZ, 16. 6. 2025 in 29. 1. 2026 na UL FRI ter 23.9.2026 na IJS).

Pri projektu do sedaj nismo zaznali odstopanj od projektne prijave in delovnega načrta. Delo poteka po načrtu. Sprotne fluktuacije kadrov vnašamo v sistem Sicris. Finančno spremljanje projekta kaže, da se sredstva porabljajo skladno s časovnim in vsebinskim načrtom dela.

Kazalnik 7.1.:

- Letno poročilo M24 (to poročilo)

1.7.2. Aktivnost 7.2: Upravljanje etike, kakovosti, tveganj in podatkov (M1-M36)

Pripravili smo Etični kodeks dela na projektu, v sklopu katerega odredimo odobritve in nadzorujemo etiko v raziskovanju in uporabi velikih jezikovnih modelov.

Po navodilih ARIS smo izdelali načrt ravnanja z raziskovalnimi podatki, ki ga z razvojem projekta in podatkov sproti nadgrajujemo.

Na vsakem polletnem srečanju projektne skupine predstavimo novosti s področja etike in se pogovorimo o morebitnih novih etičnih dilemah in tveganjih.

Kazalniki 7.2:

- Kazalnik v tem obdobju ni bil predviden

1.7.3. Aktivnost 7.3: Komunikacija, diseminacija in eksploatacija (M1-M36)

Projekt je usmerjen v izboljšanje velikih jezikovnih modelov za slovenščino ter razvoj inovativnih pristopov k raziskovanju na področjih digitalne humanistike, kot so zgodovina, folkloristika, leksikografija in pravo. Ključni cilji projekta so oblikovanje kakovostnih podatkovnih virov, razvoj aplikacij in orodij na osnovi velikih jezikovnih modelov in prispevanje k novemu znanju na področju digitalne humanistike.

Načrt diseminacije in komunikacije projekta temelji na treh osrednjih stebrih: 1) izobraževanje (dvig zavedanja o rezultatih, transparentnosti in zaupanja javnosti v velike jezikovne modele), 2) sodelovanje (vključevanje raziskovalcev in javnosti, predvsem z aktivnostmi občanske znanosti), in 3) uporaba (spodbujanje rabe podatkovnih zbirk, razvitih aplikacij in orodij). Pričujoče poročilo je strukturirano glede na ciljne skupine komunikacije in diseminacije projekta.

Raziskovalna skupnost

- **Znanstveni članki:** Cilj (7 člankov/leto) smo presegli. Med septembrom 2025 in septembrom 2026 smo objavili 15 izvirnih znanstvenih člankov. Objave so med drugim v revijah *Computer Speech & Language*, *Frontiers in Artificial Intelligence*, *Journal of Machine Learning Research*, *Slovenščina 2.0*, *Slavistična revija* in druge.
 - **Konferenčni prispevki:** Cilj (7 prispevkov/leto) je bil močno presežen. Med septembrom 2025 in septembrom 2026 smo predstavili 30 konferenčnih prispevkov. Med pomembnejšimi konferencami so *Language Resources and Evaluation Conference (LREC) 2026* in *Jezikovne tehnologije in digitalna humanistika (JT-DH)*, ki smo jo tudi organizirali v sodelovanju z drugimi raziskovalnimi institucijami.
 - **Izmenjava znanja:** Aktivno sodelovanje v COST akcijah, organizacija dogodkov v sodelovanju s Slovenskim društvom za jezikovne tehnologije in vabljenja predavanj na tujih in domačih univerzah.
- Spletna prisotnost:** Redne objave o rezultatih na spletni strani in na družbenih omrežjih (Instagram, Facebook, LinkedIn Centra za jezikovne vire in tehnologije).

Študenti in izobraževalne institucije

- **Webinarji:** Izvedenih je bilo 5 webinarjev v sodelovanju s Slovenskim društvom za jezikovne tehnologije v obliki JOTA predavanj. Izvedene so bile CLASSLA delavnic doma in v tujini. Izvedli smo delavnico o razvoju osebnih umetnointeligenčnih agentov v sodelovanju s Klubom ADA – skupnost žensk v tehnologiji.

Odločevalci in oblikovalci politik

- Številni v projekt vključeni raziskovalci so vodilni na svojih področjih in pogosto nastopajo na dogodkih za širšo skupnost in ožje skupine odločevalcev. Raziskovalci so sodelovali na naslednjih panelnih diskusijah in dogodkih: 7. dan komunikatorjev javnega sektorja, srečanje Zavoda za razvoj umetne inteligence, Dnevi slovenske informatike, Ministrstvo za visoko šolstvo, znanost in inovacije, Krepitev digitalnih znanj in spretnosti javnih uslužbencev in 5. nacionalna konferenca o umetni inteligenci.

Partnerji v industriji

- Izvedba predavanja na internem dogodka podjetja DELL.
- Redne objave na LinkedInu, namenjene predvsem deležnikom v industriji.
- Sodelovanje s podjetjem NVIDIA pri iniciativi Sovereign AI.
- Sodelovanje z industrijskimi partnerji sicer poteka pretežno preko projektov Slovenska tovarna umetne inteligence (SLAIF) in Kompetenčni center za umetno inteligenco (KCUI), pri katerih sodeluje precej vključenih raziskovalcev iz UL in IJS.

Splošna javnost

- **Družbena omrežja:** Cilj (1 objava/teden) smo presegli – v enem letu smo na LinkedIn, Instagram in Facebook objavili 90 vsebin o projektu in projektnih aktivnostih ter rezultatih.
- **Mediji:** Projekt je bil predstavljen v 16 medijskih objavah (članki, radijski intervjuji) v oddajah medijskih hiš RTV Slovenija in POP TV. Presežen je cilj 5 poslanih medijskih dopisov.
- **Občanska znanost:** Redna vabila občanskim znanstvenikom in znanstvenicam, da prispevajo podatke o govorjeni slovenščini. Pomemben dosežek je uspeh projekta *Od občanske znanosti do digitalne slovarske baze* na razpisu ARIS. V okviru projekta smo leksikografsko pregledali in validirali sopomenke in protipomenke, ki jih je prispevala javnost in jih vključili v Digitalno slovarsko bazo za slovenščino.

Kazalniki 7.3:

- Kazalnik v tem obdobju ni bil predviden

2. Najpomembnejši dosežki projektne skupine v drugih dvanajstih mesecih trajanja projekta

2.1. Dosežki na raziskovalnem področju

2.1.1. Izziv 1 Izboljšanje metod velikih jezikovnih modelov z jezikovnimi viri in razvoj slikovno-jezikovnih modelov

YADAV, Pooja, HARJULE, Priyanka, AGARWAL, Basant, ROBNIK ŠIKONJA, Marko. Assessing reliability of BERT-based models on question answering tasks. *Journal of experimental & theoretical artificial intelligence*. [Print ed.]. 2026, vol. , no. , str. 1-21, graf. prikazi. DOI: [10.1080/0952813X.2026.2716084](https://doi.org/10.1080/0952813X.2026.2716084). [COBISS.SI-ID [288277507](https://cobiss.si/id/288277507)]

ULČAR, Matej, ŽAGAR, Aleš, ARMENDARIZ, Carlos S., REPAR, Andraž, POLLAK, Senja, PURVER, Matthew, ROBNIK ŠIKONJA, Marko. Mono- and cross-lingual evaluation of representation language models on less-resourced languages. *Computer speech & language*. Jan. 2026, vol. 95, [article no.] 101852, 1-29 str., ilustr. DOI: [10.1016/j.csl.2025.101852](https://doi.org/10.1016/j.csl.2025.101852). [COBISS.SI-ID [241622275](https://cobiss.si/id/241622275)]

MARTINC, Matej, VREŠ, Domen. A large-scale instruction-tuning dataset and models for Slovenian vision-language tasks. V: PIPERIDIS, Stelios (ur.). *The Fifteenth Language Resources and Evaluation Conference (LREC 2026) : main conference proceedings : 13-15 May, 2026, [Palma, Mallorca]*. [Palma de Mallorca]: ELRA Language Resources Association, cop. 2026. Str. 9420-9433, ilustr. LREC proceedings. DOI: [10.63317/2e3jf6e7tcoh](https://doi.org/10.63317/2e3jf6e7tcoh). [COBISS.SI-ID [285727235](https://cobiss.si/285727235)]

ČIBEJ, Jaka. Metoda strojne gradnje arhetipskih besedotvornih dreves na podlagi Slovenskega oblikoslovnega leksikona Sloleks. V: KERN, Boris (ur.). *Stopenjsko besedotvorje v slovanskih jezikih*. 1. e-izd. Ljubljana: Založba ZRC, ZRC SAZU, 2026. Str. 517-531, ilustr. <https://omp.zrc-sazu.si/zalozba/catalog/view/2256/9105/3295>. [COBISS.SI-ID [271709955](https://cobiss.si/271709955)]

KNEZ, Timotej, ŽITNIK, Slavko. Improving Slovene language models for lexicographic question answering through continued pretraining and instruction fine-tuning. V: HINRICHS, Erhard (ur.). *LREC 2026 : Workshop on Structured Linguistic Data and Evaluation (SLiDE) : workshop proceedings : May 11, 2026, [Palma, Mallorca]*. [S. l.]: ELRA Language Resources Association, cop. 2026. Str. 114-123, ilustr. <https://www.slide-workshop.org/book.pdf#page=128>. [COBISS.SI-ID [277877507](https://cobiss.si/277877507)].

2.1.2. Izziv 2 Veliki jezikovni modeli za jezikoslovje in upravljanje znanja

ŠKVORC, Tadej, ROBNIK ŠIKONJA, Marko. Slovene word-sense disambiguation and novelty detection with dictionary examples. *Data & knowledge engineering*. [Print ed.]. Nov. 2026, vol. 166, article no. 102638, str. 1-15, ilustr. DOI: [10.1016/j.datak.2026.102638](https://doi.org/10.1016/j.datak.2026.102638). [COBISS.SI-ID [290903811](https://cobiss.si/290903811)]

Špela Arhar Holdt, Tina Munda, Nejc Robida, Živa Štebljaj, Polona Gantar, v tisku: Od avtentičnih popravkov do generiranih napak: učni podatki za slovnično pregledovanje slovenščine. Obdobja 2026.

Štebljaj, Živa, Munda, Tina, Arhar Holdt, Špela, Robida, Nejc, Gantar, Polona, & Robnik-Šikonja, Marko (2026). Strojno podprta kontrola kakovosti pri pripravi učnih podatkov za slovnično pregledovanje. Conference on Language Technologies and Digital Humanities (JT-DH-2026), Ljubljana, Slovenia. <https://doi.org/10.5281/zenodo.22766141>

TERČON, Luka, DOBROVOLJC ZOR, Kaja, KLEMEN, Matej, ARČON, Tjaša, ROBNIK ŠIKONJA, Marko. *Corpus-grounded evaluation dataset for grammatical question answering GramQA 1.0*. Ljubljana: University of Ljubljana, Centre for Language Resources and Technologies, 2026. 1 spletni vir. CLARIN.SI data & tools. <https://www.clarin.si/repository/xmlui/handle/11356/2086?show=full&locale-attribute=sl>. [COBISS.SI-ID [286008579](https://cobiss.si/286008579)]

ARČON, Tjaša, KLEMEN, Matej, ROBNIK ŠIKONJA, Marko, DOBROVOLJC ZOR, Kaja, TERČON, Luka. *A multilingual benchmark for evaluating metalinguistic knowledge WALS-Bench 1.0*. Ljubljana: University of Ljubljana, Centre for Language Resources and Technologies, 2026. 1 spletni vir. CLARIN.SI data & tools. <https://www.clarin.si/repository/xmlui/handle/11356/2083>. [COBISS.SI-ID [286053635](https://cobiss.si/286053635)]

2.1.3. Izziv 3 Veliki jezikovni modeli za govorjeni jezik

VERDONIK, Darinka, VIDINIČ, Jasna. Dialoška dejanja v zasebni govorni interakciji = Dialogue acts in private spoken interaction. *Slavistična revija : časopis za jezikoslovje in literarne vede*. [Tiskana izd.]. 2026, letn. 74, št. 1, str. 77–96. DOI: [10.57589/srl.v74i1.4295](https://doi.org/10.57589/srl.v74i1.4295). [COBISS.SI-ID [277004803](https://cobiss.si/277004803)], VERDONIK, Darinka, RUPNIK, Peter, LJUBEŠIČ, Nikola. Disfluencies in public and private speech. *Language and speech*. 2026, vol. 69, issue 3, str. 778-804, ilustr. DOI: [10.1177/00238309251380518](https://doi.org/10.1177/00238309251380518). [COBISS.SI-ID [256844291](https://cobiss.si/256844291)]

ŽGANK, Andrej, DONAJ, Gregor, KOLARIČ, Urh, SEREINIG, Uši, KOREN ZWITTER, Tatjana, BOTO, S., ZWITTER, Sabina, VIDINIČ, Jasna, VERDONIK, Darinka. Developing Zila: a spoken language resource for the endangered slovenian Gail Valley dialect. V: PIPERIDIS, Stelios (ur.). *The Fifteenth Language Resources and Evaluation Conference (LREC 2026) : main conference proceedings : 13-15 May, 2026, [Palma, Mallorca]*. [Palma de Mallorca]: ELRA Language Resources Association, cop. 2026. Str. 3325-3332, ilustr. LREC proceedings. <http://www.lrec-conf.org/proceedings/lrec2026/LREC-2026.pdf>. [COBISS.SI-ID 279378435]

VERDONIK, Darinka, BIZJAK, Andreja, DONAJ, Gregor. Govorjena slovenščina: structured conversational data collection through online user. V: GRISOT, Cristina (ur.), KONTINO, Thalassia (ur.). *CLARIN annual conference proceedings 2025 : 30 September – 2 October 2025 Vienna, Austria*. [S. l.: s. n.], 2025. Str. 6-10, ilustr. CLARIN Annual Conference Proceedings. DOI: <https://doi.org/10.3384/ecp222.1597> https://www.clarin.eu/sites/default/files/CLARIN2025_ConferenceProceedings.pdf. [COBISS.SI-ID 265365763]

Zgank, A., Kolaric, U., & Kacic, Z. (2026, September). How Robust are Modern Automatic Speech Recognizers to Domain and Temporal Mismatch? A Study on the Less-Resourced Slovenian Language. In *International Conference on Speech and Computer* (pp. 55-67). Cham: Springer Nature Switzerland. (COBISS VNOS V PRIPRAVI).

Andrej Zgank, Gregor Donaj, Urh Kolaric, Usi Sereinig, Tatjana Koren-Zwitter, Sanja Boto, Sabina Zwitter-Grilc, Jasna Vidinic & Darinka Verdonik Korpus Zila: Govorni vir za raziskave in razvoj avtomatskega razpoznavanja za ziljsko narečje. Konferenca JTDH 2026, Ljubljana, Slovenija. (COBISS VNOS V PRIPRAVI).

D. Vlaj, M. Kos, A. Žgank: Comparative Analysis of GMM, HMM, and LSTM-Based Models for Speaker Gender and Age Classification using Identical Acoustic Features. Konferenca IEEE ELEKTRO 2026, Žilina, Slovaška. V tisku.

VERDONIK, Darinka. Narečni korpus Zila – od posnetka do analize. V: *Govorjena slovenščina*. Maribor: Inštitut za elektrotehniko in telekomunikacije. Avg. 2026, 2 str. <https://govorjena-slovenscina.um.si/narecni-korpus-zila-od-posnetka-do-analize/>. [COBISS.SI-ID 290320899]

VERDONIK, Darinka, RUPNIK, Peter, VIDINIČ, Jasna, LJUBEŠIČ, Nikola. *Corpus of spoken Slovenian ROG-Dialog 1.0*. Ljubljana: Centre for Language Resources and Technologies, University of Ljubljana, 2025. 1 spletni vir. CLARIN.SI data & tools. <http://hdl.handle.net/11356/2073>. [COBISS.SI-ID 260584707]

2.1.4. Izziv 4 Napredne tehnologije za digitalno humanistiko

MACHIDON, Octavian-Mihai, REJEC, Jasmina, VREŠ, Domen, MACHIDON, Alina Luminita. Large language models for OCR in cultural heritage: a comparative study on Slovene folkloristic texts. *International journal on document analysis and recognition*. 2026, vol., no., str. 1-15, ilustr. DOI: [10.1007/s10032-026-00612-7](https://doi.org/10.1007/s10032-026-00612-7). [COBISS.SI-ID 287945475]

MALENŠEK, Miha, ZAVRŠNIK, Aleš, KRAJNC, Saša, KRIŽNAR, Primož, BAJEC, Marko, ŽITNIK, Slavko. Towards contradiction detection in legal texts : corpus preparation and contradiction extraction. *Slovenščina 2.0 : empirične, aplikativne in interdisciplinarne raziskave*. 2025, letn. 13, št. 2, str. 179-209, ilustr. DOI: 10.4312/slo2.0.2025.2.179-209. [COBISS.SI-ID 266721539].

Chatterjee N, Koloski B, Doucet A, Pollak S and Purver M (2026) Evaluating fine-tuned, embedding-based, and zero-shot models for aspect-based sentiment analysis in South Slavic news. *Front. Artif. Intell.* 9:1844418. doi: 10.3389/frai.2026.1844418

KOLOSKI, Boshko, POLLAK, Senja, NAVIGLI, Roberto, ŠKRLJ, Blaž. FuDoBa : fusing document and knowledge graph based representations with Bayesian optimisation. *Machine learning*. 6 Mar. 2026, vol. 115, article no. 61, str. 1-39, ilustr. DOI: [10.1007/s10994-026-07008-y](https://doi.org/10.1007/s10994-026-07008-y). [COBISS.SI-ID 271609091]

Marko Pranjic, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2026. Tracking semantic change in Slovene: A novel dataset and optimal transport-based distance. Accepted in PeerJ.

Šmajdek, U., & Bohak, C. (2025). NERVIS : an interactive system for graph-based exploration and editing of named entities. HCI SI Koper 2025 : Proceedings of the 10th Human-Computer Interaction Slovenia Conference : University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Koper, Slovenia, 13 October 2025, 147–156. <https://doi.org/10.26493/978-961-293-559-7.14>

Šmajdek, U., & Bohak, C. (2026). EntityWeaver: Visual Exploration and Curation of Named-Entity Relationships in Document Collections. <https://arxiv.org/abs/2609.37329>

Uroš Šmajdek, Tina Munda & Matija Marolt. (2026). Low-Resource Historical Named Entity Recognition in the Era of Large Language Models. JT-DH 2026

Munda, T., Dobranić, F., Šmajdek, U., Pejić, O., Bohak, C., Gorjanc, V., & Fišer, D. (2026). Benchmarking LLMs for aspect-based sentiment classification in Slovene historical periodicals. LREC 2026 : Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026 : Workshop Proceedings, 103–113. <https://doi.org/10.63317/22hvcbc23rts>

Vojko Gorjanc, Filip Dobranić, Uroš Šmajdek, Oliver Pejić, Tina Munda & Darja Fišer. (2026). Combining Graph Visualisation and Critical Discourse Analysis in the Study of Slovene Historical Newspapers. JT-DH 2026.

ATANASOVSKA, Elena, KOLOSKI, Boshko, KOCEV, Dragi. Evaluation of knowledge graph construction methods on the stroke domain. V: KOPRINSKA, Irena (ur.), MENDES-MOREIRA, João (ur.), BRANCO, Paula (ur.). *Machine learning and principles and practice of knowledge discovery in databases. Part 1 : International Workshops of ECML PKDD 2025, Porto, Portugal, September 15–19, 2025 : revised selected papers*. Cham: Springer, cop. 2026. Str. 360-372, ilustr. *Communications in computer and information science (Internet)*, 2839. DOI: 10.1007/978-3-032-19096-3_23. [COBISS.SI-ID 282082051].

ATANASOVSKA, Elena. Evaluation of knowledge graph construction methods on the stroke domain : bachelor's thesis : undergraduate university study programme, first cycle Computer and Information Science. Ljubljana: [E. Atanasovska], 2026. 1 spletni vir (1 datoteka PDF (62 str.)), ilustr. Repozitorij Univerze v Ljubljani – RUL. [COBISS.SI-ID 275585283]

ATANASOSKA, Elena, KOLOSKI, Boshko, KOCEV, Dragi. Evaluation of knowledge graph construction methods for the stroke domain. V: International conference AI for science 2025 : Ljubljana, Slovenia, 22.09.2025-26.09.2025 : including DS-2025, the SMASHing conference (4 tracks) and 4 additional tracks. Ljubljana: University of Ljubljana, Faculty of computer and information science, 2025. Str. 41-42. https://ai4science.si/wp-content/uploads/2025/09/AI4Science_Book_of_Abstracts-4.pdf. [COBISS.SI-ID 251639555]

MALENŠEK, Miha, KRAJNC, Saša. Zaznava protislovij v slovenskih pravnih besedilih : priprava korpusa za strojno učenje. Pravna praksa : PP. [Tiskana izd.]. 2. jul. 2026, leto 45, št. 26, str. 14-16, tabela, portreta. [COBISS.SI-ID 283369219]

2.1.5. Izziv 5 Izbrani izzivi digitalne humanstike

KRIŽNAR, Primož, MALENŠEK, Miha. Sistem za preverjanje skladnosti pravnih dokumentov : visoko tvegani sistem umetne inteligence?. Pravna praksa : PP. [Tiskana izd.]. 18. jun. 2026, leto 45, št. 25, str. 10-12, z avtorj. sl. [COBISS.SI-ID 282043139]

Babnik, Jan; Tratnik, Polona (2026): Political Memory as Agonistic Practice on Social Media: Semio-Somatic Memory, Multimodality, and Affordances Theorized Through the Digital Circulation of the Slogan "Smrt fašizmu, svoboda narodu". *Annales, Series Historia et Sociologia*, 36, 2, 205–226. <https://doi.org/10.19233/ASHS.2026.12>

Horvat, Marjan (2026): From Memory Regimes to Discursive Modes: A Theory-Driven Framework for Analysing Cultural Memory in Hybrid Public Spheres. *Annales, Series Historia et Sociologia*, 36, 2, 163–182. <https://doi.org/10.19233/ASHS.2026.10>

Horvat, Marjan; Koražija, Jure (2026): Conflict, Antagonistic Tone, and Deliberative Quality in Online Memory Debates: Europe Day and the Fall of the Berlin Wall on Twitter/X. *Annales, Series Historia et Sociologia*, 36, 2, 247–266. <https://doi.org/10.19233/ASHS.2026.14>

Horvat, Marjan; Koražija, Jure; Babnik, Jan; Škvorc, Tadej; Darovec, Darko; Oman, Žiga; Lampe, Urška; Ergaver, Angelika; Robnik-Šikonja, Marko (2026): Mapping Contested Cultural Memory: An LLM-Supported Approach to Analysing Narrative Structures, Discursive Modes and Discourse Functions. *Annales, Series Historia et Sociologia*, 36, 2, 183–204. <https://doi.org/10.19233/ASHS.2026.11>

Lampe, Urška; Horvat, Marjan; Koražija, Jure; Ergaver, Angelika; Darovec, Darko (2026): Agonistic Engagement in Memory Politics: Media Arenas, Normative Orientations, and Debates on the *Giorno del Ricordo* in Italy and Slovenia. *Annales, Series Historia et Sociologia*, 36, 2, 227–246. <https://doi.org/10.19233/ASHS.2026.13>

Machidon, Octavian M.; Rejec, Jasmina; Vreš, Domen; Machidon, Alina L. (2026): Large Language Models for OCR in Cultural Heritage: A Comparative Study on Slovene Folkloristic Texts. *International Journal on Document Analysis and Recognition (IJDAR)*. <https://doi.org/10.1007/s10032-026-00612-7>

Škvorc, Tadej; Horvat, Marjan; Koražija, Jure; Robnik-Šikonja, Marko (2026): Towards Future Artificial Intelligence Agents for Improved Political Discourse Quality with Large Language Models. *Annales, Series Historia et Sociologia*, 36, 2, 267–288. <https://doi.org/10.19233/ASHS.2026.15>

Arčon, Tjaša; Robnik-Šikonja, Marko; Tratnik, Polona. (2026): Large Language Models for Folktale Type Automation Based on Motifs: Cinderella Case Study. *Fabula*.

Darovec, Darko; Koražija, Jure (2026): Peace as Strong as Oak: Ritual Law, Vendetta, and Mediation in Renaissance Europe and the Americas. Preprint.

2.1.6. Izziv 6 Evalvacija in razumevanje velikih jezikovnih modelov

DEBEVEC, Luka, MARTINC, Matej. 2026. Lost in Translation or Found in Ideology? A Robust Statistical Assessment of Political Bias and Steerability in Slovenian Large Language Models. 15th Conference on Language Technologies and Digital Humanities (JT-DH-2026), held in Ljubljana, Slovenia, in September 2026. <https://zenodo.org/records/22765800>

KUZMAN PUNGERŠEK, Taja, RUPNIK, Peter, LJUBEŠIĆ, Nikola, et al. Supercharging agenda setting research : The ParlaCAP dataset of 28 European parliaments and a scalable multilingual LLM-based classification. V: *The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026 : Workshop Proceedings : 16 May 2026*. [Palma de Mallorca]: ELRA Language Resources Association, cop. 2026. Str. 94-110, ilustr. DOI: [10.63317/4t2m5y2szdw8](https://doi.org/10.63317/4t2m5y2szdw8). [COBISS.SI-ID [288964611](https://cobiss.si/288964611)]

KUZMAN PUNGERŠEK, Taja, RUPNIK, Peter, SUCHOMEL, Vit, LJUBEŠIĆ, Nikola. The growing gains and pains of iterative web corpora crawling : insights from south Slavic CLASSLA-web 2.0 corpora. V: PIPERIDIS, Stelios (ur.). *The Fifteenth Language Resources and Evaluation Conference (LREC 2026) : main conference proceedings : 13-15 May, 2026, [Palma, Mallorca]*. [Palma de Mallorca]: ELRA Language Resources Association, cop. 2026. Str. 7545-7555, ilustr. LREC proceedings. DOI: [10.63317/2p8hh2c5z2wv](https://doi.org/10.63317/2p8hh2c5z2wv). [COBISS.SI-ID [288949251](https://cobiss.si/288949251)]

LJUBEŠIĆ, Nikola, RUPNIK, Peter, PORUPSKI, Ivan, KUZMAN PUNGERŠEK, Taja. ParlaSpeech 3.0 : richly annotated spoken parliamentary corpora of Croatian, Czech, Polish, and Serbian. V: PIPERIDIS, Stelios (ur.). *The Fifteenth Language Resources and Evaluation Conference (LREC 2026) : main conference proceedings : 13-15 May, 2026, [Palma, Mallorca]*. [Palma de Mallorca]: ELRA Language Resources Association, cop. 2026. Str. 5677-5688, ilustr. LREC proceedings. DOI: [10.63317/3b7dnbjr75es](https://doi.org/10.63317/3b7dnbjr75es). [COBISS.SI-ID [288941315](https://cobiss.si/288941315)]

VINTAR, Špela, KUZMAN PUNGERŠEK, Taja, BRGLEZ, Mojca, LJUBEŠIĆ, Nikola. Charting the European LLM benchmarking landscape : a new taxonomy and registry. V: MONTEJO-RÁEZ, Arturo (ur.), GRISOT, Cristina (ur.), BLOCHOWIAK, Joanna (ur.). *LREC 2026 : Shaping multilingual, multimodal AI for the social sciences and humanities (LLMs4SSH) @ LREC 2026 : workshop proceedings*. [Palma de Mallorca]: ELRA Language Resources Association, cop. 2026. Str. 205-217, ilustr. DOI: [10.63317/4kixe3c9zmdc](https://doi.org/10.63317/4kixe3c9zmdc). [COBISS.SI-ID [288956419](https://cobiss.si/288956419)]

KUZMAN PUNGERŠEK, Taja, RUPNIK, Peter, PORUPSKI, Ivan, DINIĆ, Vuk, LJUBEŠIĆ, Nikola. State of the art in text classification for south slavic languages : fine-tuning or prompting?. V: MONTEJO-RÁEZ, Arturo (ur.), GRISOT, Cristina (ur.), BLOCHOWIAK, Joanna (ur.). *LREC 2026 : Shaping multilingual, multimodal AI for the social sciences and humanities (LLMs4SSH) @ LREC 2026 : workshop proceedings*. [Palma de Mallorca]: ELRA Language Resources Association, cop. 2026. Str. 1-13, ilustr. DOI: [10.63317/2fudy99w2taz](https://doi.org/10.63317/2fudy99w2taz). [COBISS.SI-ID [288954883](https://cobiss.si/288954883)]

MIOK, Kristian, KLEMEN, Matej, ŠKRLJ, Blaž, ROBNIK ŠIKONJA, Marko. Challenges in explaining pretrained clinical text classifiers. V: KOPRINSKA, Irena (ur.), MENDES-MOREIRA, João (ur.), BRANCO, Paula (ur.). *Machine learning and principles and practice of knowledge discovery in databases. Part 4 : International Workshops of ECML PKDD 2025, Porto, Portugal, September 15–19, 2025 : revised selected papers*. Cham: Springer, cop. 2026. Str. 314–322, graf. prikazi. Communications in computer and information science (Internet), 2842. DOI: [10.1007/978-3-032-19105-2_22](https://doi.org/10.1007/978-3-032-19105-2_22). [COBISS.SI-ID [278437123](https://cobiss.si/278437123)]

MIOK, Kristian, ŠKRLJ, Blaž, ZAHARIE, Daniela, ROBNIK ŠIKONJA, Marko. TT-XAI: trustworthy clinical text explanations via keyword distillation and LLM reasoning. V: KOPRINSKA, Irena (ur.), MENDES-MOREIRA, João (ur.), BRANCO, Paula (ur.). *Machine learning and principles and practice of knowledge discovery in*

databases. Part 4 : International Workshops of ECML PKDD 2025, Porto, Portugal, September 15–19, 2025 : revised selected papers. Cham: Springer, cop. 2026. Str. 499–513, ilustr. Communications in computer and information science (Internet), 2842 DOI: [10.1007/978-3-032-19105-2_35](https://doi.org/10.1007/978-3-032-19105-2_35). [COBISS.SI-ID [278451715](https://cobiss.si/278451715)]

KUZMAN PUNGERŠEK, Taja, LJUBEŠIĆ, Nikola. *Multilingual training dataset for CAP policy topic classification ParlaCAP-train*. Ljubljana: Centre for Language Resources and Technologies, University of Ljubljana, 2026. 1 spletni vir. CLARIN.SI data & tools. <https://www.clarin.si/repository/xmlui/handle/11356/2093?show=full>. [COBISS.SI-ID [286004483](https://cobiss.si/286004483)]

KUZMAN PUNGERŠEK, Taja, RUPNIK, Peter, LJUBEŠIĆ, Nikola. *South Slavic web corpus collection CLASSLA-web 2.0*. Ljubljana: Jožef Stefan Institute, 2026. 1 spletni vir. CLARIN.SI data & tools. <https://www.clarin.si/repository/xmlui/handle/11356/2079>. [COBISS.SI-ID [272257027](https://cobiss.si/272257027)]

2.2. Drugi pomembni rezultati

Raziskovalna skupina je dosegla številne druge pomembne rezultate, kot so prostodostopni veliki jezikovni modeli, številne podatkovne množice, prostodostopna programska koda in mednarodna ter nacionalna sodelovanja. Rezultati na področju komunikacije, diseminacije in eksploatacije so opisani v razdelku 1.7.