Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije

Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities: LLM4DH

Izroček projekta/ Project Deliverable

Letno poročilo M12 / Anual Report M12

Št. Izročka / Deliverable No.	7.1.1
Datum oddaje / Delivery date	27.10. 2025 (Verzija 1.2)
Vrsta / Nature[1]:	R
Nivo diseminacije / Disemination level[1]:	PU
Delovni Sklop / Work package	WP7 Management and Dissemination T7.1 Project coordination & monitoring
Avtor, Vodilni partner / Author, Lead partner	Marko Robnik-Šikonja, UL FRI
Sodelujoči partnerji / Contributing partners	Špela Arhar Holdt, Marko Bajec, Ciril Bohak, Slavko Žitnik, Tjaša Arčon, Gašper Jelovčan, Tinca Lukan, Maja Zupančič Justin, Alina-Luminita, Machidoni , Mihai Octavian Machidoni, Tadej Škvorc, Aleš Žagar, Žiga Lesar, Matija Marolt, Alenka Kavčič, Miha Malenšek, Timotej Knez UL FRI Senja Pollak, Simon Krek, Nikola Ljubešić, Matej Martinc, Marko Pranjić, Matthew Purver, Jaya Caporusso IJS Darja Fišer, Andrej Pančur, Jakob Lenardič INZ Polona Tratnik, IRRIS Aleš Završnik, Primož Križnar, IK Kaja Dobrovoljc, Špela Vintar, Mojca Brglez, UL FF Simon Dobrišek, UL FE Darinka Verdonik, Andrej Žgank, UM FERI

[1]Nature: R = Report, P = Prototype, D = Demonstrator, O = Other (specify)
[1]Dissemination level PU = Public ; CO = Confidential, only for members of the consortium (including ARIS)

Projekt financira ARIS – Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije, št projekta GC-0002

Informacijska tabela projekta / Project inforamtion 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)finanfiranje Gravitacije, št. 5100-70/2024-11 z dne 29.5.2024.
Koodrinator projekta / Project coordinator	Marko Robnik-Šikonja
Trajanje projekta / Project duration	1.10.2025 – 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 prpojekta/ 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 / Document history			
Datum / Date	Verzija	Avtor / Author	Komentar / Comment
27.8.2025	1.0	Maja Zupančič Justin	Postavitev zasnove poročila
september 1015	1.1	Vsi vodje DS in DN	dopolnitve
27.10.2025	1.2	Marko Robnik Šikonja	Zaključna redakcija

Kazalo vsebine / Table of content

1. Rezultati in dosežki projekta: realizacija programa dela v prvih dvanajstih mesecih trajanja projekta

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.

Te zakladnice informacij veliki jezikovni modeli še ne uporabljajo ustrezno, vendar bi lahko zmanjšale halucinacije, izboljšale njihovo jezikovno znanje v zapletenih situacijah in jezikih, podprtimi z manj viri, ter izboljšale prilagoditve jezikovnih modelov za določene pomembne naloge, kot sta zdravorazumsko sklepanje in sklepanje v naravnem jeziku.

Delovni sklop 1 (DS1) vsebuje tri naloge, ki naslavljajo

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

Za izboljšanje slovenskih velikih jezikovnih modelov (VJM) smo raziskovali možnosti uporabe dodatnih podatkov v različnih fazah učenja modelov. Podprli smo dva tipa izboljšav: (1) v fazi predučenja, kjer smo model opremili z dodatnim slovarskim znanjem, in (2) v fazi prilagajanja modela za odgovarjanje na vprašanja, kjer smo model dodatno učili odgovarjati na leksikografska vprašanja.

Najprej smo pripravili obsežno zbirko slovarskih podatkov iz več virov: Digitalne slovarske baze (DSB), slovarja Bridge, podatkovne zbirke za razpoznavo pomena besede, slovenskega slovarja sopomenk in odprte baze WordNet (Čibej idr. 2023). Vse podatke smo združili v enotno strukturirano obliko, ki vključuje besedne oblike, pomene, definicije, kolokacije, sinonime in primere rabe. Zbrali smo približno 356.000 besed, od tega večino (269.940) samostalnikov. Na podlagi te zbirke smo avtomatsko ustvarili besedilni korpus, ki vsebuje opise besed in je namenjen dodatnemu vnaprejšnjemu učenju slovenskih VJM. Korpus je trenutno objavljen na spletni strani projekta in obsega 73.531.189 besed.

Drugi del učnih podatkov je bil pripravljen za prilagajanje modela za odgovarjanje na vprašanja. V ta namen smo iz zbirke besed pripravili korpus vprašanj in odgovorov. Vprašanja smo oblikovali na več načinov:

- avtomatsko generirana vprašanja o oblikah, pomenih, sopomenkah, kolokacijah in kontekstih rabe,
- vprašanja o besedni analizi stavkov, sestavljena na podlagi podatkovne zbirke ssj500k,
- ročno oblikovana vprašanja in odgovori, zbrani na dogodku *promptathon* (25 vprašanj) ter v Jezikovni svetovalnici (3.698 unikatnih vprašanj).

Za povečanje raznolikosti smo pri avtomatsko ustvarjenih primerih uporabili model GPT-4.1, ki je vprašanja preoblikoval v bolj raznolike oblike.

Na podlagi teh podatkov smo pripravili začetni izboljšani VJM. Kot osnovo smo vzeli *GaMS-Instruct* model, ki smo ga dodatno doučili na pripravljeni zbirki vprašanj in odgovorov. Model smo ovrednotili z metodo primerjave, pri kateri je »sodniški VJM« (GPT-4.1) ocenjeval kakovost odgovorov. V prvem scenariju je ocenjeval semantično bližino odgovora glede na referenčni odgovor, v drugem pa primerjal

dva odgovora po slovnični pravilnosti. V obeh primerih je na testnem delu naše zbirke vprašanj dodatno učeni model dosledno presegel tako osnovni *GaMS-Instruct* model kot tudi *GaMS-Instruct-nemotron* model.

Kazalniki 1.1:

- Podatkovne zbirke DSB in OSWN, pripravljene za učenje (M6, Simon Krek)
- Začetni izboljšani veliki jezikovni model (M12, Simon Krek)
- Končni izboljšani veliki jezikovni modeli (M24).

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

Velike količine večjezičnih in medjezikovnih jezikovnih podatkov v več jezikih ponujajo priložnost za izboljšanje VJM z nadaljnjim vnaprejšnjim učenjem in prilagajanjem posebnim nalogam. Z večjezičnim znanjem izboljšani večjezični VJM bodo izboljšali medjezikovni prenos, kar je še posebej pomemb za jezike z manj viri, kot je slovenščina.

Z uporabo večjezičnih digitalnih slovarskih podatkovnih zbirk, kot sta Wiktionary in slovarji, smo začeli razvoj metodologije za vnos znanja v VJM. Metodologija bo vsebovala gradnjo grafov znanja in direktno generiranje sintetičnih besedil. Izluščene grafe znanja in generirana besedila smo zaenkrat preizkusili na problemu izboljšanja strojnih prevajalnikov pri prevajanju figurativnega jezika.

Kazalniki 1.2: Grafi znanja in podatkovne zbirke neobdelanih besedil (M18). Začetni izboljšani veliki jezikovni model (M24). Končni izboljšani veliki jezikovni model (M30).

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

Cilji: 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).

Rezultati: V prvem letu projekta smo uspešno dosegli prvi ključni mejnik (M12) z objavo obsežne podatkovne zbirke za učenje slikovno-jezikovnih modelov za slovenski jezik. Podatkovna zbirka, poimenovana SLO-VLM-IT-Dataset 1.0, je ključnega pomena za nadaljnji razvoj na področju multimodalnega strojnega učenja za jezike z manj viri. Zbirka, ki je javno dostopna v repozitoriju CLARIN.SI, vsebuje skupno 1.128.228 primerov, pri čemer je vsak učni primer sestavljen iz para slike in pripadajočega opisnega besedila, kar omogoča modelu učenje povezav med vizualno in tekstovno modaliteto. Za zagotavljanje raznolikosti podatkov je sestavljena iz strojno prevedene uveljavljene angleške zbirke Llava_v1.5_mix665k ter iz podatkov, specifičnih za slovensko okolje. Ti so bili generirani s pomočjo modela Gemini 1.5 Pro na podlagi slik in člankov iz slovenske Wikipedije ter vodilnih slovenskih novičarskih portalov, kot so RTV SLO, Siol in 24ur.

Podatkovna zbirka je strukturirana tako, da vključuje širok spekter nalog, ključnih za učenje naprednih slikovno-jezikovnih zmožnosti. Med drugim vsebuje primere za odgovarjanje na vprašanja na podlagi prepoznave besedil na slikah (OCR), generiranje opisov slik, odgovarjanje na vprašanja z enobesednimi ali večizbirnimi odgovori na podlagi slik iz zbirke COCO ter povezovanje opisov z deli slik, kjer model določa koordinate za določen opis ali opiše področje, določeno s koordinatami.

Na podlagi te novonastale podatkovne zbirke je bil že narejen pomemben korak k doseganju drugega cilja projekta (M24). Razvita je bila prva različica slovenskega slikovno-jezikovnega modela, ki temelji na odprtokodnem modelu google/gemma-3-4b-it. Model je bil dodatno učen z metodo nadzorovanega prilagajanja (ang. supervised fine-tuning) in je javno dostopen za preizkušanje in nadaljnje raziskave na platformi Hugging Face.

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)

V zadnjih tridesetih letih so leksikografi, ki se ukvarjajo z leksikalnimi opisi, poskušali uporabiti različne avtomatizirane postopke za analizo jezikovnih podatkov in izdelavo leksikografskih opisov (Atkins in Rundell 2008, Gantar et al. 2016). Zaradi najnovejšega razvoja na področju generativne umetne inteligence je prišlo do poskusov uporabe novih orodij v leksikografske namene (Lew 2023; Rees idr. 2023; Jakubíček in Rundell 2023). Vendar se je po prvih poskusih izkazalo, da obstaja precejšen razmak med kakovostjo leksikografskih vsebin, ki jih modeli generirajo v angleščini, in tistimi, ki jih generirajo v drugih jezikih, zlasti v jezikih z manj viri ali tisti, ki so v modelih manj zastopani (de Schryver, 2024).

Digitalna slovarska zbirka za slovenščino (DSB) na različnih ravneh jezikovnega opisa izboljšujemo z uporabo modelov, izdelanih v aktivnosti 1.1. 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. Rezultat bo bistveno izboljšana DSB, ki jo bomo nato uporabili za izboljšanje modelov v aktivnosti 1.1. Različice DSB so na voljo kot javno dostopne podatkovne zbirke in prek API-ja z odprtim dostopom.

Kazalniki 2.1: Digitalna slovarska baza z generiranimi leksikografskimi podatki – prva različica (M24), digitalna slovarska baza z generiranimi leksikografskimi podatki – končna različica (M36)

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

Cilji: Podatkovne množice sintetičnih jezikovnih napak (M12). Veliki jezikovni modeli za preverjanje slovnice (M18). Evalvacijske množice za preverjanje avtentične slovnice (M24).

Rezultati: V prvem letu projekta smo razvijali metodologijo za strojno generiranje povedi, ki vsebujejo raznolike in avtentične jezikovne napake v slovenščini, skupaj z jezikovnim popravkom in kategorijo (vrsto) tega popravka. V prvem delu smo sistematično preizkusili različne strategije za generiranje tovrstnih podatkov z velikimi jezikovnimi modeli. Testirali smo več modelov (GPT-4o, DeepSeek 2.5, GaMS-9B-Instruct, Grok 3) in zanje pripravili različne tipe vhodnih podatkov ter oblikovali pozive, ki temeljijo na empiričnih podatkih iz korpusov Šolar 3.0 in Gigafida 2.0.

Kljub prilagajanju vhodov in pozivov so bili rezultati nezadovoljivi: generirane napake so bile pogosto pretesno kopirane po vhodnih primerih ali pa so bile nerealistične oziroma neavtentične. Za uspešnejše strojno generiranje smo oblikovali ročno revidirano podatkovno množico DASSLE 1.0 (*Dataset of Authentic and Synthetic Slovene Language Errors,* Kazalnik 2.2.1, dostopen na repozitoriju CLARIN.SI). Množica obsega 7.385 korpusno pridobljenih ali sintetično ustvarjenih povedi, umeščenih v 5 jezikovnih ravnin (črkovanje, pravopis, oblikoslovje, besedišče, skladnja) in 128 specifičnih vrst, pri čemer vsaka poved vsebuje en, homogeno opredeljen in popravljen problem. Podatki, ki smo jih predstavili na konferenci (Jelovčan et al. 2025), bodo osnova za nadaljnje delo – učenje modelov za slovnično pregledovanje slovenščine.

Sodelovali smo pri skupni nalogi Multi-GEC 2025, ki se posveča razvoju in evalvaciji novih metod strojnega popravljanja za različne evropske jezike. Naše sodelovanje je zagotovilo, da je v učnih in evalvacijskih podatkih skupne naloge zastopana tudi slovenščina. Rezultate smo popisali v prispevku za znanstveno revijo (Masciolini et al. 2025).

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

Cilji: Veliki jezikovni modeli z izboljšanim slovničnim znanjem (M12). Podatkovna množica za evalvacijo slovničnega znanja velikih jezikovnih modelov (M18). Večjezične in medjezične slovnične analize (M36).

Rezultati: V prvem letu projekta smo uspešno razvili prvo različico na velikih jezikovnih modelih temelječega sistema za napredno analizo slovnično razčlenjenih besedilnih zbirk (Kazalnik 2.3.1, M12). Sistem z delovnim imenom UD-LLM temelji na podatkih večjezične zbirke Universal Dependencies (UD) in deluje kot prilagojena večfazna agentna arhitektura, ki združuje analizo poizvedb, generiranje kode in s poizvedovanjem obogateno tehniko generiranja besedil (RAG). Ta kot vhod sprejme uporabniško slovnično poizvedbo (npr. »Kakšen je vrstni red pridevnika in samostalnika v slovenščini?«), nato pa na podlagi podatkov iz zbirke UD uporabniku poda večplastni odgovor. VJM najprej zgenerira programsko kodo v jeziku python, ki se izvede na celotni zbirki UD za izbrani jezik, rezultati pa se združijo v skupno distribucijo in dopolnijo s povzetkom najpogostejšega odgovora.

Razviti sistem smo testirali na 20 izbranih slovničnih vprašanjih na temo besednega reda za 179 jezikov, vključenih v zbirko UD v2.16. Kot jezikovni model smo uporabili model GPT-5. Merili smo klasifikacijsko točnost najpogostejšega odgovora, preciznost in priklic veljavnih možnosti za izbrani jezik in ustreznost izračunane distribucije na podlagi Hellingerjeve razdalje. Primerjava med razvitim sistemom in jezikovnim modelom brez prilagoditev je za večino problemov pokazala jasne izboljšave, na nekaj problemih pa smo ugotovili omejitve trenutnega sistema, ki jih aktivno razrešujemo. Koda razvitega sistema je odprto dostopna na platformi GitHub [1].

Poleg zgoraj navedenih aktivnosti, povezanih z realizacijo kazalnikov, smo v okviru projekta razvili tudi več povezanih virov in tehnologij za slovnično analizo slovenskega jezika, kot je nova različica govorne drevesnice SST [2], z imenskimi entitetami nadgrajena različica pisne drevesnice SSJ [3], novi modeli strojnega označevalnika CLASSLA-Stanza [4] ter nadgrajena različica orodja STARK [5]. Organizirali smo tudi mednarodno multikonferenco SyntaxFest 2025, ki je potekala na Pravni fakulteti Univerze v Ljubljani med 26. In 29. avgustom 2025 [6].

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, pritrjevanje, 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)

Cilji: Zasnovati spletni vmesnik za zbiranje podatkov o pogovornem govoru, pri čemer bodo upoštevani ekonomski, etični in pravni vidiki ter poenostavljeno in avtomatizirano zbiranje metapodatkov (M4). Zbrani in ročno transkribirani bodo pogovorni podatki za govorni učni korpus (5 ur) (M7). Zbrani in transkribirani bodo podatki za SloBench ASR korpus v obsegu 1 ure ter dodani k obstoječim 3 uram (M25). Izdana bo avdio baza konverzacijskega govora (M36).

Rezultati: Vzpostavljen je bil portal za zbiranje govornih posnetkov na daljavo z angažiranjem občanskih raziskovalcev. Izdelana so bila navodila za snemanje (COBISS.SI-ID 224971267). S pomočjo najetih pravnih strokovnjakov iz APHAIA so bile postavljene pravne podlage za zbiranje podatkov in upravljanje osebnih podatkov, ki se zbirajo v bazi. Definiran je bil seznam metapodatkov o posnetkih in govorcih, ki se zbirajo ob posnetkih. Zbiranje podatkov je bilo digitalizirano, tako da poteka prek spletnega vmesnika. Vzpostavljena je bila podatkovna baza, v kateri se urejajo in arhivirajo vsi zajeti podatki. Prek portala je bilo zbranih 5 ur posnetkov zasebne interakcije iz vseh slovenskih statističnih regij, ki predstavljajo osnovo za učni govorni korpus ROG-dialog. Posnetki so bili transkribirani v pogovornem in standardiziranem načinu skladno s pravili transkribiranja. Za zagotavljanje točnosti transkripcij je vsaka transkripcija po prvem zapisu šla še skozi tri korake pregleda. Prav tako je bila prek portala zbrana 1 ura posnetkov za nadgradnjo evalvacijske baze SloBench ASR. Portal je bil obogaten s poljudnimi vsebinami o značilnostih govora (COBISS.SI-ID 226075651; COBISS.SI-ID 246993155; COBISS.SI-ID 246992643; COBISS.SI-ID 246993923; COBISS.SI-ID 246993667; COBISS.SI-ID 246994179). Predstavitev portala strokovni javnosti bo potekala 1. 10. 2025 na Clarin Annual Conference na Dunaju.

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

Cilji: Ročno označena govorna dejanja in sentiment v pogovorih v obsegu 5 ur govora (M10). Modeli za označevanje govornih dejanj in sentimenta v govorjeni slovenščini (M30).

Rezultati: Posnetki in transkripcije korpusa ROG-dialog, zbrani v T3.1, so bili označeni z dialoškimi dejanji in sentimentom. Opravljen je bil pregled shem za označevanje dialoških dejanj s poudarkom na generičnih shemah: DAMLS, SWBD-DAMSL, AMI, DART, DIT++, ISO 24617-2. Kot najbolj aktualna, razširjena in dobro dokumentirana je bila izbrana shema ISO 24617-2. Na podlagi te sheme so bile najprej natančno opredeljene enote označevanja. Sledili smo široko sprejeti enoti označevanja v govoru, C-unit (Biber et al., 1999: Longman Grammar of Spoken and Written English). Opredelili smo jo z upoštevanjem prozodičnih lastnosti govora in z minimalističnim pristopom. Po izvedeni segmentaciji je sledilo dodajanje oznak za dimenzije in komunikacijske funkcije, kot jih definira standard ISO 24617-2. V drugem koraku je sledil pregled in izbira sheme za označevanje sentimenta. Opredelili smo 6-stopenjsko označevalno lestvico od pozitivnega prek nevtralnega do negativnega sentimenta. Enota označevanja je bila v osnovi C-unit, kot je bila opredeljena že za dialoška dejanja, vendar je bilo omogočeno združevanje več enot v eno glede na to, kaj je minimalna enota govora, ki ji je mogoče določiti sentiment. Rezultat je ročno označen korpus govorjene interakcije ROG-dialog v obsegu 5 ur. Vse oznake so ročno dodane. Za zagotavljanje kvalitete so vse dodane oznake tudi pregledane in popravljene s strani nadzorujočega označevalca. Postopek objave korpusa v repozitoriju CLARIN.SI je v teku.

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

Cilji: Zbirka slovenskih govornih posnetkov iz javno dostopnih virov namenjena avtomatskemu razpoznavanju govora (min. 300 ur) (M36).

Rezultati: Cilj delovnega sklopa je zasnovati in preveriti možnosti učinkovitega zbiranja slovenskih govornih posnetkov, ki služijo za samo-nadzorovano oziroma nenadzorovano učenje avtomatskega razpoznavanja govora. Takšen tip zbirke se uporablja brez ročno tvorjenih transkripcij, kar bistveno poenostavi in poceni njegovo izdelavo. Izvorni nabor posnetkov se praviloma pridobi iz javno dostopnih virov na internetu, kar pomeni, da je lahko vprašljiva njihova kakovost. Zato smo najprej izvedli pregled metrik za vrednotenje govornega in zvočnega signala (PEAQ, PESQ, POLQA, NISQA, VISQOL itd.). Kot potencialno primerne smo identificirali metrike brez reference, ki temeljijo na pristopih globokega učenja. Za pripravo akustične metrike in modeliranja avtomatskega razpoznavalnika govora se uporabljajo značilke. Tako smo pripravili analizo značilk, vezanih na avtomatsko razpoznavanje govora in sorodne govorne tehnologije. Identificirali smo sledeče možne pristope: MFCC, PLP, spektrogram.

LPC, valjčna transformacija, ZCR, signalne značilnost. Nato smo pristopili k analizi razpoložljivih slovenskih govornih virov na internetu, kjer smo prednost dali uporabniško tvorjenim vsebinam, ki so na voljo preko pretočnih in multimedijskih storitev. Izbrali smo prvi omejen nabor posnetkov, ki bo služil za izdelavo preliminarne analize. Za namen preliminarne analize smo začeli s postavitvijo izhodiščnega avtomatskega razpoznavalnika govora za slovenski jezik.

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

Izvedli smo obsežno študijo o plitvi integraciji modelov avtomatske razpoznave govora in velikih jezikovnih modelov ter izmerili in primerjali učinkovitost teh metod. V študijo smo vključili osem metod plitve integracije: ponovno ocenjevanje hipotez (rescoring) z deskriptivnim modelom, ponovno ocenjevanje hipotez z generativnimi modeli, popravljanje napak v hipotezah z generativnimi modeli z dodatnimi primeri in brez njih (zero- in one-shot generative error correction), aktivacijski pristop k popravljanju napak, izbor najboljše hipoteze, učeno mapiranje hipoteza-transkript (pristop vključuje prilagajanje generativnih modelov) in arhitekturo SpeechLLM, kjer se veliki jezikovni model nauči pisanja transkripta iz rezultatov razpoznavalnika.

V sklopu teh metod plitve integracije smo testirali 8 velikih jezikovnih modelov (SloBERTa, GaMS-9B, EuroLLM-9b, gemma-2-9b, gemma-3-12b, gemma-3-27b, Lamma-3.1-8b in SlovenianGPT) in njihove variacije za sledenje navodilom (instruct), kjer so bile potrebne. Ker se večina plitvih integracij izvaja nad hipotezami modela za avtomatsko prepoznavanje govora (automatic speech recognition, ASR), smo v študiji raziskali tudi kako količina videnih hipotez vpliva na uspešnost metod; te so tako v eksperimentih videle od 5 do 100 hipotez.

Kombinatorično glede na uporabljen model, pristop in količino videnih hipotez smo izvedli 97 eksperimentov nad množicami Artur, CommonVoice, Fleurs, GVL, Sofes in VoxPopuli.

V primerjavi z osnovnim modelom z N-gramskim jezikovnim modelom slovenskega jezika (metrika Word Error Rate – WER), so se najbolje izkazale metode ponovnega ocenjevanja hipotez, ki so imele nižji WER kot osnovni model. Ostale metode so dosegle večinoma primerljive rezultate brez izboljšav, z nekaj izboljšav videnih tudi v pristopu mapiranja hipoteza-transkript. Opazili smo tudi, da uporabljen ukazni poziv (prompt) precej vpliva na uspešnost modela; variacije ukaznih pozivov je zato potrebno še dodatno raziskati.

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.

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

Pripravili smo prvi nabor zgodovinskih parlamentarnih dokumentov, z ročno označenimi imenskimi entitetami (na GitHub, bo v CLARIN). Na podlagi teh označenih podatkov smo razvili interaktivno vizualizacijo interakcijskih grafov (D4.1, M12), ki prikazuje omrežja ljudi, organizacij, krajev in drugih entitet ter njihove medsebojne povezave. Vizualizacija omogoča raziskovanje gostote povezav, pojavnosti entitet ter izvajanje popravkov, tako na entitetah, kot na povezavah med njimi. V povezavi s tem smo pripravili članek *NERVIS: An Interactive System for Graph-Based Exploration and Editing of Named Entities*, ki čaka na objavo.

Prav tako smo preverili učinkovitost obstoječih modelov za prepoznavanje imenskih entitet na izbranih zgodovinskih korpusih. Pokazalo se je, da se ti modeli na zgodovinskih podatkih obnesejo precej slabše kot na modernih.

Vzporedno smo pripravili razširjen nabor smernic za označevanje imenskih entitet v zgodovinskih zapisih. Za podporo nadaljnjemu razvoju sistema smo začeli z gradnjo podatkovne zbirke zgodovinskih slovenskih oseb in krajev, ki bo služila kot referenčna osnova za postopke povezovanja entitet.

Poleg tega smo objavili še pregledno poglavje v knjigi o integraciji velikih jezikovnih modelov in grafov znanja ter prispevek na delavnici o velikih jezikovnih modelih za presojanje tripletov v grafih znanja na specifičnem področju kemije.

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.

Ker se je aktivnost začela v šestem mesecu (M6), je bil v prvem letu poudarek na pripravi eksperimentalnega okolja in začetnih testiranjih. Obstoječi slovenski VJM (GaMS) smo preizkusili na zlatem standardu za zaznavanje semantičnih sprememb (Pranjić et al., 2024), vendar rezultati zaenkrat ne presegajo naključnega ugibanja.

Obstoječi pristop, ki smo ga razvili pred začetkom projekta (Pranjić et al., 2024) in temelji na vložitvah SloBERTa, smo dodatno analizirali in prispevek poslali v recenzijo v revijo PeerJ.

Vzpostavili smo tudi alternativni pristop, ki temelji na vložitvah SentenceBERT, pri katerem so vložitve povezane v graf na podlagi semantične podobnosti. Graf je nato segmentiran v gruče, ki združujejo močno povezane pomenske sklope vsake besede. Pristop je bil uporabljen na korpusu slovenskih periodičnih publikacij (sPeriodika 1.0), s fokusom na temi revščine in bolezni. Analizirali smo besede, kot so *ubog*, *revež*, *sirota*, *siromak*, in njihovo rabo skozi čas. Semantične skupine so bile vizualizirane na časovni premici, njihova interpretacija pa podprta z izluščenimi ključnimi besedami iz stavkov. V sodelovanju z Univerzo v Edinburgu smo metodo preizkusili tudi na angleškem korpusu SHE s področja medicine. Čeprav je pristop še v fazi razvoja, kaže potencial za zgodovinsko analizo semantičnih sprememb in rab skozi čas ter omogoča uporabo na zelo obsežnih besedilnih korpusih.

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

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

Rezultati: Aktivnost je v začetni fazi (začetek M10), zato je bil poudarek na razvoju teoretičnega okvira, ki je ključen za nadaljnje delo. Rezultati tega začetnega dela so vidni v dveh sprejetih znanstvenih objavah, ki postavljata temelje za prihajajoče raziskave. Na konferenci AI for Science je bil sprejet abstrakt z naslovom "Teaching Vision-Language Models Semiotics: Toward a Socio-Semiotic Framework for Multimodal AI" (Martinc, Babnik, 2025), na delavnici "AI Methods for Research of Folkloristic Narratives" pa povzetek "Considering Modes: Semiotics and Multimodal AI" (Babnik, Martinc, 2025). Obe objavi predstavljata nov semiotični okvir, ki bo služil kot osnova za učenje slikovno-jezikovnih modelov prepoznavanja globljega pomena v vizualnih gradivih. Predstavljen je bil tudi osnutek učnega pristopa, ki se osredotoča predvsem na izboljšanje sposobnosti slovenskega slikovno-jezikovnega modela za iskanje in analizo fotografij ter na prilagoditev teh modelov za rabo v digitalni humanistiki.

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 bi avtomatsko prepoznaval vsebinska neskladja, zato predstavlja pomemben korak k učinkovitejši pravni analizi in razvoju naprednih pravnih informacijskih sistemov in kvalitetnejši pravni nomotehniki.

V podrobnem pregledu področja smo pripravili članek za revijo Slovenščina 2.0, Towards contradiction detection in legal texts: corpus preparation and contradiction extraction. Dela zadnjih let raziskav se osredotočajo predvsem na povzemanje, napovedovanje sodb ali na nalogo odgovarjanja na vprašanja. Naši izbrani nalogi, zaznavi protislovij, je še najbližje iskanje pravnih argumentov (Legal Argument Mining) (Poudyal et al., 2020; Habernal et al. 2023), kar pa se kljub temu precej razlikuje od zaznave protislovij.

Korpus pripravljen v sklopu 1.5.3 smo uporabili za predučenje novega deskriptivnega modela slovenskega jezika, z arhitekturo ModernBERT (Answerdotai, 2024), ki podpira daljše dokumente, značilne za pravno domeno. Prvo domensko prilagojeno predrazličico modela, PravniBERT, smo že pripravili, glavnina predučenja pa je še v teku, saj je za domensko prilagojen model potrebno pripraviti tudi splošen slovenski model.

S predrazličico modela PravniBERT smo izvedli uvodno raziskavo detekcije protislovij v pravnih besedilih, ki je temeljila na Stvarnopravnem zakoniku. Pripravili smo umetno množico za sklepanje v naravnem jeziku (Natural Language Inference, NLI), ročno pregledano s strani pravnih strokovnjakov ter nov pristop k predučenju modela za stavčne vložitve, temelječi na predrazličici modela PravniBERT. Na nalogi iskanja pravnih kontradikcij za dani člen Stvarnopravnega zakonika (SPZ), je pilotni model dosegel natančnost najboljših treh (top 3) 83.6%, kar pomeni, da se je pravilno protislovje pojavilo v prvih treh rezultatih, ki jih model vrne. Kljub umetnim podatkom je rezultat dober in je motiviral večjo označevalno kampanjo z ročnim označevanjem, ki temelji na Kazenskem zakoniku, Obligacijskem zakoniku, Stvarnopravnem zakoniku in Zakonu o varstvu osebnih podatkov. Objavo nastale učne množice načrtujemo do konca leta.

Dosedanje rezultate in delo smo predstavili tudi na konferenci Razumeti pravo 2025, na Pravni fakulteti Univerze v Ljubljani.

1.5. Izziv 5: Izbrani izzivi digitalne humanstike

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

V preteklem letu smo uspešno izvedli niz ključnih aktivnosti, usmerjenih v razvoj metodološke osnove za kritično analizo zgodovinskega diskurza s pomočjo sodobnih jezikovnih tehnologij.

Opravili smo sistematični pregled literature, ki zajema metodološke pristope, orodja in tehnične rešitve za uporabo VJM v digitalni humanistiki s posebnim fokusom na njihovi uporabi za analizo zgodovinskih besedil. Zbrana literatura je objavljena v javni zbirki Zotero in raziskovalkam ter raziskovalcem ponuja vstopno točko v raziskovanje grafov imenskih entitet (T5.1-O1) in diahrone študije konceptov na podlagi zgodovinskih virov (T5.1-O3).

Obravnavali smo izzive, povezane z arhaičnimi jezikovnimi oblikami in specifikami zgodovinskega konteksta, ki otežujejo neposredno uporabo sodobnih orodij za obdelavo naravnega jezika (natural language processing, NLP) na tiskanih zgodovinskih virih. Med najpogostejšimi posebnostmi zgodovinskega tiska najdemo pomanjkanje velikih začetnic, nestandardno črkovanje, tipografske posebnosti, kot so npr. veliki razmiki med črkami namesto odebeljenega ali poševnega tiska itn. V sodelovanju z relevantnimi področnimi strokovnjaki (jezikovni tehnologi, jezikoslovci in zgodovinarji) smo opravili evalvacijo treh modelov (CLASSLA-Stanza v2, SloNER, XLMR Slavic) za prepoznavanje imenskih entitet (NER) s primerjavo posebej za ta namen ročno označenega vzorca zgodovinskih virov,

ki vsebuje izvoda časopisov ter zapisnik seje Kranjskega deželnega zbora. Na podlagi rezultatov analize pripravljamo prilagoditve smernic za označevanje NER v zgodovinskih tekstih, ki bodo javno objavljene do konca 2025. Ključna omejitev obstoječih smernic za označevanje NER pri njihovi uporabi na zgodovinskih virih je njihovo zanašanje na sodobne standarde pisane slovenščine, ki, še posebej pred letom 1899, niso obstajali.

Pomemben prispevek je tudi priprava tematskega podkorpusa sPeriodika, ki zajema ključne slovenske časopise 19. in začetka 20. stoletja. Izbor virov temelji na posvetih z domenskimi strokovnjaki (zgodovinarji, sociolingvisti), kar zagotavlja relevanten in strokovno utemeljen izbor virov za analizo družbenopolitičnega diskurza na ozemlju današnje Slovenije. Ta podkorpus bomo v nadaljevanju uporabili za analizo interakcij med imenskimi entitetami s kombiniranjem metodoloških pristopov kvantitativne analize omrežij in kritične analize diskurza (T5.1-M1). Celoten korpus sPeriodika je objavljen na repozitoriju in konkordančniku CLARIN.SI, pripravljen podkorpus pa je trenutno dostopen na kot predpripravljena poizvedba na konkordančniku, njegovo ločeno objavo pa pričakujemo skupaj z objavo prispevka, ki ga raziskuje.

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

Cilj naloge T5.2 je razviti metodologijo, ki temelji na velikih jezikovnih modelih (VJM), za računalniško 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 pravljice, ljudske pripovedi in druge etnografske naracije niso zgolj brezčasne zgodbe, temveč odsevi družb, ki so jih ustvarjale in posredovale. Osrednji izziv je zato avtomatizacija analize pripovednih struktur ob preučevanju različnih različic zgodb skozi čas in kulture.

Rezultati: V prvem letu projekta smo se osredotočili na oblikovanje digitalnih korpusov, ki predstavljajo temelj za nadaljnji razvoj metodologije računalniške folkloristike z uporabo velikih jezikovnih modelov. Trenutno imamo štiri ključne zbirke: (1) etnografska besedila, (2) pravljice iz zbirke ZRC SAZU, (3) revija *Ciciban* in (4) *Kmetijske in rokodelske novice (KRN)*. Prvi, drugi in četrti korpus so že oblikovani, tretji (Ciciban), ki je kompleksno multimodalen, je v zaključni fazi. Tipične težave pri optičnem prepoznavanju (optical character recognition, OCR) zgodovinskega gradiva in starejših revij, kot je *Ciciban*, vključujejo zgodovinsko ortografijo, raznolike tipografije, ročno pisavo, kompleksne postavitve strani, kombinacijo slike in besede, nečitljiv ali poškodovan tisk, ligature in posebne znake, dialektalne oblike ter slabo kakovost prebranih dokumentov. Težave zgodovinske ortografije, raznolikih tipografij, rokopisov, dialektov, kompleksnih postavitev in multimodalnih gradiv smo premostili z uporabo KrakenOCR in prilagojenih modelov v eScriptorium, polavtomatiziranimi predprocesnimi postopki (OCR refinement, normalizacija), pilotnimi testi z VJM-ji ter razvojem prioritetnega sistema za oblikovanje besedilnih in slikovnih korpusov, pri čemer sodelujemo tudi z raziskovalci, ki razvijajo vizualne modele (T4.3), saj tako gradimo cevovod za nadaljnjo uporabo pri analizi motivov, vrednot in pripovednih struktur.

Pomemben rezultat predstavlja izvedba mednarodne delavnice *AI Methods for the Research of Folkloristic Narratives*, ki je povezala raziskovalce iz folkloristike, digitalne humanistike, računalništva in kulturne zgodovine. Potekala je 12. junija 2025 na UL FRI v soorganizaciji z Inštitutom IRRIS. Uvodno plenarno predavanje je imel Marko Robnik Šikonja (*Large Language Models for Analysis of Complex Phenomena*), poseben gost pa je bil Timothey Tangherlini, vodilni raziskovalec digitalne folkloristike v svetovnem merilu, ki je imel plenarno predavanje *AI Archives: Leveraging Multilingual LLMs for Folklore Archives*. Člani našega raziskovalnega sklopa so predstavili več referatov: Tjaša Arčon, Marko Robnik Šikonja in Polona Tratnik (*Motif Detection Using LLMs: The Cinderella Case Study*), Jasmina Rejec (*Computational Detection of Moral Values in Slovenian Animal Folktales Using GPT-4 and Lexicon-Based Methods*), Marjan Horvat, Jure Koražija in Polona Tratnik (*Modeling Deliberative Values in Narrative Culture Using Large Language Models*), Jan Babnik in Polona Tratnik (*The Dragon-Slayer's Narrative: Structural Kinship and Discursive Divergence*), Darko Darovec, Angelika Ergaver, Jure Koražija, Veronika Kos in Žiga Oman (*The Earliest Accounts of the Americas and the Customary Conflict Resolution*

Analyzed by OpenAI and ChatGPT) ter Jan Babnik in Matej Martinc (*Considering Modes: Semiotics and Multimodal AI*). Poleg njih so sodelovali tudi drugi raziskovalci, med drugim Sándor Darányi, Thomas van Erven in Judith Veld, kar je omogočilo široko mednarodno razpravo in utrdilo metodološki ter interdisciplinarni okvir projekta.

Jasmina Rejec je predstavila prispevek *"Animal Folk Tales as Vehicles of Moral Values and Social Critique"* na 14. Mednarodni konferenci mladih folkloristov (14th International Conference of Young Folklorists) v Tartu v Estoniji.

V okviru diseminacije smo pripravili tri izvirne znanstvene članke, ki smo jih poslali v recenzijo v revije:

- Arčon, Robnik Šikonja, Tratnik: *Large Language Models for Folktale Type Automation Based on Motifs: Cinderella Case Study*
- Machidon, Rejec, Vreš, Machidon: *Large Language Models for OCR in Cultural Heritage: A Comparative Study on Slovene Folkloristic Texts*
- Horvat, Koražija, Babnik, Škvorc, Robnik-Šikonja: *Cultural Memory from Antagonism to Deliberation in (Social) Media: AI Approach*

Razvoj korpusov, metodologije in diseminacijskih aktivnosti kaže, da projekt uspešno napreduje proti osrednjemu izročku – novi metodologiji za digitalno folkloristiko, predvideni za marec 2026.

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

Za predučenje ali domensko prilagajanje jezikovnih modelov je potrebna velika količina teksta. Za potrebe pravnega sistema, podprtega s sistemom za s pridobivanjem obogatenim generiranje (retrieval augmented generation, RAG), smo pripravili največji korpus pravnega slovenskega jezika, požet iz spletnih virov pravnih besedil (PISRS, USRS, Sodna praksa, Državni list). Prečiščen korpus vsebuje 1 milijardo žetonov (angl. tokens) in ohranja vse metapodatke, ki jih spletni viri ponujajo. Uporabljen je za aktivnosti in prototipe v 1.4.4.

Proces je vključil implementacijo namenskih spletnih pajkov v kombinaciji z uporabo ponujenih API klicev na spletiščih, ki so to ponujala. Nabrane dokumente smo prečistili in segmentirali glede na vire, ter, kjer so prisotni, na posamezne člene. Korpus zaobjema vse objave omenjenih spletišč od 1. 1. 1991 do 31. 12. 2024.

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 prošenj in faux pasov (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 (Sravanthi 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 pri projektu zastavili več ciljev, s katerimi zapolnjujemo vrzeli na področju evalvacijskih množic za slovenščino, in sicer za področje pragmatičnega pomena in figurativnega jezika.

Izdelali smo dve podatkovni množici, in sicer:

- SloPragEval: Podatkovna množica je slovenska različica množice MultiPragEval (Park in dr. 2024), ki vsebuje 300 primerov za preskušanje pragmatičnega jezikovnega razumevanja. Primeri temeljijo na Griceovih maksimah sodelovalne komunikacije, in sicer je za vsako od štirih maksim vključenih 60 primerov, pri katerih pride do kršitve maksime, nadaljnjih 60 primerov pa je dobesednih (brez kršitve). Podatkovno množico smo ročno prevedli in večkrat revidirali, številne primere pa smo tudi kulturno prilagodili. Poleg tega smo prek spletne kampanje pridobili človeške rešitve, ki jih je prispevalo prek 80 sodelujočih. Povprečna človeška natančnost je 0,85, kar omogoča primerjavo jezikovnih modelov s to referenčno vrednostjo.
- SloPragMega: Podatkovna množica je prevedena in prilagojena različica treh nalog iz množice PragMega (Hu in dr. 2023), in sicer sklopov za metaforo, ironijo in humor. Naloge so zasnovane v obliki vprašanj z več možnimi odgovori, skupno število nalog pa je 105. Naloge za razumevanje metafore in ironije so oblikovane tako, da je v odgovorih treba poiskati pravo interpretacijo konteksta, pri nalogah za humor pa je ponujenih več možnih nadaljevanj, od katerih je treba izbrati najbolj humorno možnost. Tudi to množico smo prevedli ročno, jo večkrat revidirali in nekaj primerov povsem spremenili, da ustrezajo slovenskemu jezikovnemu in kulturnemu okolju.

Obe podatkovni množici sta objavljeni na evalvacijski platformi SloBench. Množica SloPragEval je bila predstavljena na konferenci AI4Science v sklopu AI for Digital Humanities, naslov prispevka: Brglez, M., Vintar, Š. (2025) SloPragEval: Creating the First Pragmatics Understanding Benchmark for Slovene.

V okviru tega projektnega sklopa smo raziskovali tudi asociativno vedenje velikih jezikovnih modelov in primerjali njihovo ustvarjalnost pri tvorjenju odzivov na čustvene besede. Raziskava je bila predstavljena na konferenci International Conference of Computational Creativity 2025: Vintar, Š., Javoršek, J. J. (2025) The truth is no diaper: Human and AI-generated associations to emotional words. Proceedings of ICCC25.

Kazalniki 6.1: Primerjalni test metafor, ironije in sarkazma v slovenščini (M12, Špela Vintar).

- SloPragEval,
- SloPragMega,
- Primerjalni test razlage pragmatičnega in asociativnega vedenja (M24).

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.

Cilj: Govorna množica (4 ure) z anotacijami govornega dejanja in sentimenta. Večreferenčna naloga avtomatske razpoznave govora in naloga procesiranja dialoga in sentimenta v govoru.

Rezultati: Prispevali smo k izbiri podatkov v aktivnostih 3.1 in 3.2, da bi zagotovili, da bodo podatki, ki bodo dodani v testno množico, ustrezne kakovosti in primerne zahtevnosti. Izvedli smo predhodne analize z govorom, označenim s sentimentom, da bi bolje razumeli zahtevnost naloge prepoznavanja sentimenta v govoru, kar je nujno za vzpostavitev uravnoteženega merila.

Pripravili smo dve preliminarni govorni testni množici podatkov z mednarodno dimenzijo. Prva je namenjena prepoznavanju zapolnjenih premorov in je na voljo v slovenščini, hrvaščini, srbščini, češčini in poljščini (COBISS.SI-ID 249955587). Druga testna podatkovna množica je namenjena napovedovanju primarnega besednega naglasa, testni podatki pa so na voljo v slovenščini, hrvaščini, čakavskem narečju in srbščini (COBISS.SI-ID 249836035).

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

Množica za evalvacijo pristranskosti Equity evaluation corpus - EEC (Kiritchenko in Mohammad 2018) je bila prilagojena slovenščini (EEC-SL) z upoštevanjem lokalnega konteksta, relevantnega za pristranost (npr. razlike med slovenskimi imeni in imeni s področja bivše Jugoslavije). Pripravljamo metodologijo za analizo pristranosti glede na spolno in demografsko pristranost jezikovnih modelov.

Razvili smo tudi nov pristop za razpristranjevanje multimodalnih modelov na nalogi generiranja slik iz besedil. Osredotoča se na atributa spola, vere in rase. Metodologija temelji na uporabi jezikovnih modelov z verigo misli (angl. chain-of-thought). Prispevek je bil sprejet na konferenco EMNLP v kategoriji Findings.

Kazalniki 6.3: Detekcija pristranskosti za slovenščino (M24). Pristop odpravljanja pristranskosti za velike jezikovne modele (M30). Analiza detekcije pristranskosti v govorjenem jeziku (M36).

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 še zdaleč ni dovolj. Za naslovitev problema pomanjkanja transparentnosti velikih jezikovnih modelov bomo razvili splošno metodologijo, ki jo bomo uporabili na specifičnih zahtevnih področjih, kot so zdravorazumsko sklepanje, sklepanje v naravnem jeziku, parafraziranje in računalniška folkloristika.

Kazalnik 6.4: Nova metodologija za razlage na osnovi znanja za velike jezikovne modele (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 s spremljanjem 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. V

prvem letu projekta smo izvedli dve takšni srečanji (4. 11. 2024 na INZ in 16. 6. 2025 na UL FRI). Naslednje srečanje je načrtovano 29. 1. 2026 na UL FRI.

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 poročila M12 (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 odrejamo 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.

Kazalniki 7.2: Načrt ravnanja z raziskovalnimi podatki (M6, Marko Robnik Šikonja). Aktivnosti za spremljanje etike, kakovosti in tveganj (M7 Marko Robnik Šikonja).

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

Projekt Veliki jezikovni modeli za digitalno humanistiko (LLM4DH) 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. Objavljenih je 6 člankov, 2 pa sta v recenziji. Objave so med drugim v *IEEE Access, Political Research Exchange, Expert Systems with Applications* in *Computer Speech and Language*.
- **Konferenčni prispevki:** Cilj (7 prispevkov/leto) je bil močno presežen, do septembra 2025 smo predstavili 27 prispevkov. Med pomembnejšimi konferencami so so *SyntaxFest 2025* in *AI 4 Science*. Naši raziskovalci so bili namreč vključeni tudi v organizacijo in programske odbore teh dveh konferenc.
- **Izmenjava znanja:** Aktivno sodelovanje v COST akcijah, organizacija dogodkov v sodelovanju s Slovenskim društvom za jezikovne tehnologije ter izvedba mednarodne delavnice *AI Methods for Research in Folkloristic Narratives*.
  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:** Izvedeni 3 webinarji v sodelovanju s Slovenskim društvom za jezikovne tehnologije in drugimi organizacijami. Izvedba CLASSLA delavnic doma in v tujini. Na portalu VideoLectures sta objavljena posnetka dveh webinarjev.

- **Gostujoča predavanja:** Cilj 3 gostujočih predavanj smo presegli, saj smo izvedli 4 gostujoča predavanja na domačih in tujih institucijah.

Odločevalci in oblikovalci politik

- Sodelovanje raziskovalcev na panelni diskusiji o dostopnosti jezika in vlogi umetne inteligence pri zagotavljanju enakopravnega dostopa. Cilj je bil eno tovrstno sodelovanje tekom projekta in ta cilj smo dosegli v prvem letu trajanja projekta.

Partnerji v industriji

- Sodelovanje na panelni diskusiji v okviru evropske iniciative European Language Data Space, ki spodbuja izmenjavo podatkov med akademijo in industrijo.
  Redne objave na LinkedInu, namenjene predvsem deležnikom v industriji.

Splošna javnost

- **Družbena omrežja:** Cilj (1 objava/teden) smo presegli – v enem letu smo na LinkedIn, Instagram in Facebook objavili 91 vsebin o projektu in projektnih aktivnostih ter rezultatih.
- **Mediji:** Projekt je bil predstavljen v 6 medijskih objavah (članki, radijski intervjuji). 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 bomo leksikografsko pregledali in validirali sopomenke in protipomenke, ki jih je prispevala javnost in jih vključili v Digitalno slovarsko bazo za slovenščino.

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

2.1. Dosežki na raziskovalnem področju

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

HOSTNIK, Marko, ROBNIK ŠIKONJA, Marko. Retrieval-augmented code completion for local projects using large language models. Expert systems with applications. [Print ed.]. Nov. 2025, vol. 292, [article no.] 128596, str. 1-15, ilustr. ISSN 0957-4174. https://www.sciencedirect.com/science/article/pii/S0957417425022158, DOI: 10.1016/j.eswa.2025.128596. [COBISS.SI-ID 242180867], [Odprti dostop, JCR, SNIP, WoS, Scopus]

MASCIOLINI, Arianna, CAINES, Andrew, ARHAR HOLDT, Špela, ŽAGAR, Aleš, et al. Towards better language representation in natural language processing : a multilingual dataset for text-level grammatical error correction. International journal of learner corpus research. 2025, vol. 11, iss. 2, str. 309-335, graf. prikazi. ISSN 2215-1478. https://www.jbe-platform.com/content/journals/10.1075/ijlcr.24033.mas, DOI: 10.1075/ijlcr.24033.mas. [COBISS.SI-ID 234594051], [Odprti dostop, SNIP, WoS, Scopus]

PETRIČ, Timotej, ARHAR HOLDT, Špela, ROBNIK ŠIKONJA, Marko. Pomembnost realistične evalvacije : primer popravkov sklona in števila v slovenščini z velikim jezikovnim modelom. Slovenščina 2.0 : empirično, aplikativne in interdisciplinarne raziskave. 2024, letn. 12, št. 1, str. 106-130, ilustr. ISSN 2335-2736. https://journals.uni-lj.si/slovenscina2/article/view/14902, DOI: 10.4312/slo2.0.2024.1.106-130. [COBISS.SI-ID 227633411], [Odprti dostop, SNIP, Scopus]

LJUBEŠIĆ, Nikola, PORUPSKI, Ivan, RUPNIK, Peter. Identifying filled pauses in speech across South and West Slavic languages. V: PISKORSKI, Jakub (ur.). 10th Workshop on Slavic Natural Language Processing 2025 (Slavic NLP 2025) : co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL) : July 31, 2025. Stroudsburg: Association for Computational Linguistics, cop. 2025. Str. 1-8, ilustr. https://aclanthology.org/2025.bsnlp-1.1/, DOI: 10.18653/v1/2025.bsnlp-1.1. [COBISS.SI-ID 249955587]

KRSNIK, Luka, DOBROVOLJC, Kaja. STARK : a toolkit for dependency (sub)tree extraction and analysis. V: 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025) : August 28-29, 2025 : proceedings. Kerrville: Association for Computational Linguistics, 2025. Str. 44-51, ilustr. ISBN 979-8-89176-291-6. https://aclanthology.org/2025.tlt-1.5.pdf. [COBISS.SI-ID 251611651], [Odprti dostop]

LJUBEŠIĆ, Nikola, PORUPSKI, Ivan, RUPNIK, Peter. Identifying primary stress across related languages and dialects with transformer-based speech encoder models. Interspeech. 2025, str. 5768-5772, tabele, graf. prikazi. ISSN 2958-1796. https://www.isca-archive.org/interspeech_2025/ljubesic25_interspeech.pdf, DOI: 10.21437/Interspeech.2025-205. [COBISS.SI-ID 249836035]

ARHAR HOLDT, Špela, ANTLOGA, Špela, MUNDA, Tina, PORI, Eva, KREK, Simon. From words to action : a national initiative to overcome data scarcity for the Slovene LLM. V: ARHAR HOLDT, Špela (ur.), et al. The third workshop on resources and representations for under-resourced languages and domains (RESOURCEFUL 2025) : proceedings of the 3rd workshop = March 2, 2025. Tallinn: [University of Tartu Library], 2025. Str. 130-136, ilustr. ISBN 978-9908-53-121-2. https://aclanthology.org/2025.resourceful-1.pdf. [COBISS.SI-ID 248005123]

Izziv 2 Veliki jezikovni modeli za jezikoslovje in upravljanje znanja

MASCIOLINI, Arianna, CAINES, Andrew, ARHAR HOLDT, Špela, ŽAGAR, Aleš, et al. Towards better language representation in natural language processing: a multilingual dataset for text-level grammatical error correction. *International journal of learner corpus research*. 2025, vol. 11, iss. 2, str. 309-335, DOI: 10.1075/ijlcr.24033.mas. [COBISS.SI-ID 234594051]

PETRIČ, Timotej, ARHAR HOLDT, Špela, ROBNIK ŠIKONJA, Marko. Pomembnost realistične evalvacije: primer popravkov sklona in števila v slovenščini z velikim jezikovnim modelom. *Slovenščina 2.0 : empirične, aplikativne in interdisciplinarne raziskave*. 2024, letn. 12, št. 1, str. 106-130, DOI: 10.4312/slo2.0.2024.1.106-130. [COBISS.SI-ID 227633411]

ARHAR HOLDT, Špela, ANTLOGA, Špela, MUNDA, Tina, PORI, Eva, KREK, Simon. From words to action: a national initiative to overcome data scarcity for the Slovene LLM. V: ARHAR HOLDT, Špela (ur.), et al. *The third workshop on resources and representations for under-resourced languages and domains (RESOURCEFUL 2025): proceedings of the 3rd workshop* = March 2, 2025. Tallinn: [University of Tartu Library], 2025. Str. 130-136. https://aclanthology.org/2025.resourceful-1.pdf. [COBISS.SI-ID 248005123]

Izziv 3 Veliki jezikovni modeli za govorjeni jezik

VERDONIK, Darinka. Diskurzni označevalci. V: *Govorjena slovenščina*. Maribor: Inštitut za elektrotehniko in telekomunikacije. Avg. 2025, 2 str. https://govorjena-slovenscina.um.si/znacilnosti-govora/diskurzni-oznacevalci/, Digitalna knjižnica Univerze v Mariboru – DKUM. [COBISS.SI-ID 246994179]

VERDONIK, Darinka. Enote govora. V: *Govorjena slovenščina*. Maribor: Inštitut za elektrotehniko in telekomunikacije. Avg. 2025, 2 str. https://govorjena-slovenscina.um.si/znacilnosti-govora/enote-govora/, Digitalna knjižnica Univerze v Mariboru – DKUM. [COBISS.SI-ID 246993667]

VERDONIK, Darinka. Kako tekoč je govor v resnici?. V: *Govorjena slovenščina*. Maribor: Inštitut za elektrotehniko in telekomunikacije. Avg. 2025, 2 str. https://govorjena-slovenscina.um.si/znacilnosti-govora/kako-tekoc-je-govor-v-resnici/, Digitalna knjižnica Univerze v Mariboru – DKUM. [COBISS.SI-ID 246993923]

VERDONIK, Darinka. Snemanje govora. V: *Govorjena slovenščina*. Maribor: Inštitut za elektrotehniko in telekomunikacije. Avg. 2025, 2 str. https://govorjena-slovenscina.um.si/znacilnosti-govora/zakaj-govor/snemanje-govora/, Digitalna knjižnica Univerze v Mariboru – DKUM. [COBISS.SI-ID 246992643]

VERDONIK, Darinka. Zapisovanje govora. V: *Govorjena slovenščina*. Maribor: Inštitut za elektrotehniko in telekomunikacije. Avg. 2025, 2 str. https://govorjena-slovenscina.um.si/znacilnosti-

govora/zapisovanje-govora/, Digitalna knjižnica Univerze v Mariboru – DKUM. [COBISS.SI-ID 246993155]

VERDONIK, Darinka. Zakaj govor?. V: *Govorjena slovenščina*. Maribor: Inštitut za elektrotehniko in telekomunikacije. Jan. 2025, 2 str. https://govorjena-slovenscina.um.si/znacilnosti-govora/zakaj-govor/, Digitalna knjižnica Univerze v Mariboru – DKUM. [COBISS.SI-ID 226075651]

VERDONIK, Darinka, BIZJAK, Andreja, DONAJ, Gregor, MAKAROVIČ, Boštjan, CONTERO ALMAGRO, Cristina. *Navodila za snemanje za portal Govorjena slovenščina*. Maribor: Univerza, Fakulteta za elektrotehniko, računalništvo in informatiko, 2025. 1 spletni vir (1 PDF datoteka (16 str.)). https://govorjena-slovenscina.um.si/wp-content/uploads/Navodila-za-snemanje-za-portal-Govorjena-slovenscina.pdf, Digitalna knjižnica Univerze v Mariboru – DKUM. [COBISS.SI-ID 224971267]

Izziv 4 Napredne tehnologije za digitalno humanistiko

MACHIDON, Octavian-Mihai, MACHIDON, Alina Luminita. Comparing OCR pipelines for folkloristic text digitization. V: *DH2025 - Digital Heritage International Congress 2025 : 8-13 September 2025, Siena (Italy)*. Goslar: Eurographics Association Postfach; Graz: Institute of Visual Computing: Fraunhofer Austria, cop. 2025. Str. 1-8, ilustr. ISBN 978-3-03868-277-6. https://diglib.eg.org/items/f1e68154-ef3e-4198-b764-2a693c59489e, DOI: 10.2312/dh.20253075. [COBISS.SI-ID 256535555]

Izziv 5 Izbrani izzivi digitalne humanstike

ARČON, ROBNIK ŠIKONJA, TRATNIK: *Large Language Models for Folktale Type Automation Based on Motifs: Cinderella Case Study*

MACHIDON, REJEC, VREŠ, MACHIDON: *Large Language Models for OCR in Cultural Heritage: A Comparative Study on Slovene Folkloristic Texts*

HORVAT, KORAŽIJA, BABNIK, ŠKVORC, ROBNIK-ŠIKONJA: *Cultural Memory from Antagonism to Deliberation in (Social) Media: AI Approach*

Izziv 6 Evalvacija in razumevanje velikih jezikovnih modelov

LJUBEŠIĆ, Nikola, PORUPSKI, Ivan, RUPNIK, Peter. Identifying filled pauses in speech across South and West Slavic languages. V: PISKORSKI, Jakub (ur.). *10th Workshop on Slavic Natural Language Processing 2025 (Slavic NLP 2025) : co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL) : July 31, 2025*. Stroudsburg: Association for Computational Linguistics, cop. 2025. Str. 1-8, ilustr. https://aclanthology.org/2025.bsnlp-1.1/, DOI: 10.18653/v1/2025.bsnlp-1.1. [COBISS.SI-ID 249955587]

LJUBEŠIĆ, Nikola, PORUPSKI, Ivan, RUPNIK, Peter. Identifying primary stress across related languages and dialects with transformer-based speech encoder models. *Interspeech*. 2025, str. 5768-5772, tabele, graf. prikazi. ISSN 2958-1796. https://www.isca-archive.org/interspeech_2025/ljubesic25_interspeech.pdf, DOI: 10.21437/Interspeech.2025-205. [COBISS.SI-ID 249836035]

2.2. Drugi pomembni rezultati

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

Prva različica slikovno-jezikovnega modela dosegljiva na Hugging Face na povezavi https://huggingface.co/GaMS-Beta/gemma-3-4b-fine-tuned-slo-VLM.

Izziv 2 Veliki jezikovni modeli za jezikoslovje in upravljanje znanja

JELOVČAN, G., ROBNIK-ŠIKONJA, M., ARHAR-HOLDT, Š., & VREŠ, D. (2025). Attempt to Create Synthetic Dataset for Grammar Error Correction in Slovenian Language. In Š. Arhar Holdt, S. Pollak, & M. Robnik-Šikonja (Eds.), *Artificial Intelligence and Digital Humanities: SMASH Track of AI4Science Conference Book of Abstracts,* (pp. 23–27). Ljubljana, Slovenia.

MASCIOLINI, Arianna, CAINES, Andrew, ARHAR HOLDT, Špela, ŽAGAR, Aleš, et al. Towards better language representation in natural language processing: a multilingual dataset for text-level grammatical error correction. *International journal of learner corpus research*. 2025, vol. 11, iss. 2, str. 309-335, DOI: 10.1075/ijlcr.24033.mas. [COBISS.SI-ID 234594051]

PETRIČ, Timotej, ARHAR HOLDT, Špela, ROBNIK ŠIKONJA, Marko. Pomembnost realistične evalvacije: primer popravkov sklona in števila v slovenščini z velikim jezikovnim modelom. *Slovenščina 2.0 : empirične, aplikativne in interdisciplinarne raziskave*. 2024, letn. 12, št. 1, str. 106-130, DOI: 10.4312/slo2.0.2024.1.106-130. [COBISS.SI-ID 227633411]

ARHAR HOLDT, Špela, ANTLOGA, Špela, MUNDA, Tina, PORI, Eva, KREK, Simon. From words to action: a national initiative to overcome data scarcity for the Slovene LLM. V: ARHAR HOLDT, Špela (ur.), et al. *The third workshop on resources and representations for under-resourced languages and domains (RESOURCEFUL 2025): proceedings of the 3rd workshop* = March 2, 2025. Tallinn: [University of Tartu Library], 2025. Str. 130-136. https://aclanthology.org/2025.resourceful-1.pdf. [COBISS.SI-ID 248005123]

KLEMEN, Matej, BOŽIČ, Martin, ARHAR HOLDT, Špela, ROBNIK-ŠIKONJA, Marko. Large Language Model-based Grammatical Error Correction of Slovenian School Essays. *Sodobna pedagogika 2025 – 77(143).* V tisku.

ARHAR HOLDT, Špela; ANTLOGA, Špela; GANTAR, Polona; MUNDA, Tina; ROBIDA, Nejc; ZGONC, Matjaž, 2025, *Dataset of Authentic and Synthetic Slovene Language Errors DASSLE 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/2052.

JELOVČAN, G., ROBNIK-ŠIKONJA, M., ARHAR-HOLDT, Š., & VREŠ, D. (2025). Attempt to Create Synthetic Dataset for Grammar Error Correction in Slovenian Language. In Š. Arhar Holdt, S. Pollak, & M. Robnik-Šikonja (Eds.), *Artificial Intelligence and Digital Humanities: SMASH Track of AI4Science Conference Book of Abstracts,* (pp. 23–27). Ljubljana, Slovenia.

KLEMEN, Matej, BOŽIČ, Martin, ARHAR HOLDT, Špela, ROBNIK-ŠIKONJA, Marko. Large Language Model-based Grammatical Error Correction of Slovenian School Essays. *Sodobna pedagogika 2025 – 77(143).* V tisku.

KLEMEN, M. et al. (2025). UD-LLM: agentni sistem za večjezično slovnično analizo besedil, https://github.com/matejklemen/ud_llm Dobrovoljc, K. (2025, v objavi). Treebanking Spoken Slovenian: New Data, Models, and Lessons Learned. P*rispevki za novejšo zgodovino.*

Munda, T., & Dobrovoljc, K. (2025). Korpus Universal Named Entities za slovenščino. https://github.com/UniversalNER/UNER_Slovenian_SSJ

TERČON, L., DOBROVOLJC, K., LJUBEŠIĆ, N (2025, v objavi). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *Prispevki za novejšo zgodovino.*

KRSNIK, L., & DOBROVOLJC, K. (2025). STARK: A toolkit for dependency (sub)tree extraction and analysis. *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025).*

DOBROVOLJC, K., & TERČON, L. (Eds.). (2025). *SyntaxFest 2025: 5 Events for 1 Fest of Empirical Syntax.* Založba Univerze v Ljubljani. https://doi.org/10.4312/9789612976361

Izziv 3 Veliki jezikovni modeli za govorjeni jezik

Postopek objave korpusa ROG-dialog, 5 ur govora, v repozitoriju CLARIN.SI še poteka in zbirka še nima COBISS ID.