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

Pristopi k analizi zgodovinskega časopisja z uporabo velikih jezikovnih modelov / Approaches to Analysing Historical Newspapers Using LLMs

Št. Izročka / Deliverable No.	5.1.1
Datum oddaje / Delivery date	31.3. 2026 (Verzija 1.0)
Vrsta / Nature[1]:	R
Nivo diseminacije / Disemination level[1]:	PU
Delovni Sklop / Work package	WP5 Izbrani izzivi digitalne humanistike / WP5 Challenges of digital humanities T5.1 Veliki jezikovni modeli za hiatoriografijo / Large language models for historiography
Avtor, Vodilni partner / Author, Lead partner	Darja Fišer, INZ
Sodelujoči partnerji / Contributing partners	Ciril Bohak, UL FRI, Andrej Pančur INZ

[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**)
Opis aktivnosti 5.1.1 / Description of the Action 5.1.1	Najprej bomo z grafi imenskih entitet iz aktivnosti 4.1 raziskali odnose med osebami, kraji in organizacijami v korpusu slovenske zgodovinske periodike sPeriodika 1.0 (1771-1914). Za analizo grafov imenskih entitet bomo uporabili pristop mešanih metod, pri čemer bomo kombinirali kvantitativno mrežno analizo s kritično analizo diskurza. Raziskava se bo osredotočila na nastanek in razvoj prepletenih zgodovinskih identitet: nacionalne, jezikovne, politične, družbenoekonomske in verske. Ker se lahko odnosi med zgodovinskimi identitetami skozi čas semantično spreminjajo, bomo diahrono analizo iz aktivnosti 4.2 nato uporabili za preučevanje odnosov, predsodkov in ideologij družbenih elit v različnih časovnih obdobjih. Grafe bomo uporabili kot približke za dinamično spreminjajoče se identitete, da bi raziskali ohranjanje jezikovnih ideologij in preučili, kako so bile te ideologije zgodovinsko povezane z drugimi vidiki ustvarjanja identitet posameznikov in krajev. Raziskali bomo tudi diahrone semantične premike leksike, povezane z vzponom in padcem zgodovinskih nacionalizmov, s poudarkom na konceptih državnosti. Raziskali bomo 1) nasprotja med panslovansko, jugoslovansko in slovensko identiteto, 2) kako so bili takšni pojmi povezani s ključnimi centri moči tistega časa (npr. habsburško cesarstvo do začetka 20. stoletja) in 3) kako takšne identitete vključujejo manjše regionalne identitete. First, we will use named-entity graphs from T4.1 to explore the relationships between people, places, and organizations in sPeriodika 1.0, a corpus of Slovenian historical periodicals (1771–1914). We will apply a mixed methods approach to analyze the named entity graphs, combining quantitative network analysis with critical discourse analysis. The investigation will focus on the emergence and development of intertwined historical identities: national, language, political, socio-economic, and religious. Second, as relations between historical identities can undergo semantic shifts through time, we will use diachronic analysis from T4.2 to study the attitudes, prejudices, and ideologies of social elites in different time periods. We will use the graphs as proxies for dynamically changing identities to investigate the perpetuation of language ideologies and investigate how such ideologies were historically tied to other identity-making aspects of individuals and places. We will also investigate diachronic semantic shifts of the lexical inventory related to the rise and fall of historical nationalisms, focusing on concepts of nationhood. We will investigate 1) oppositions between Pan-Slavic, Yugoslavian, and Slovenian identities, 2) how such notions were related to the major centers of power of the time (e.g., the Habsburg Empire until the early 20th century), and 3) how such identities incorporate smaller regional ones.

Kazalo vsebine / Table of content

1. Cilji

Cilj naloge je razviti mešani pristop za historiografično analizo kolektivnih identitet v zgodovinski periodiki, ki kombinira metode oddaljenega in bližnjega branja ter že uveljavljene metode tematskega modeliranja in grafov imenskih entitet nadgradijo z določanjem sentimenta do imenskih entitet z uporabo velikih jezikovnih modelov.

2. Rezultati

Razvili smo leksikon kolektivnih identitet in testno množico za vrednotenje označevanja sentimenta do imenskih entitet (obe sta javno objavljeni in dostopni na https://dihur.si/muki/llm4dh/).

Tematsko modeliranje in analizo časopisov Slovenka, Slovenec in Slovenski narod smo izvedli z BERTopic in pripravili prispevek za PressMint-LREC2026 (Dobranić et al., v objavi).

Označevanje sentimenta do imenskih entitet pa smo preizkusili na modelu GaMS ter še na treh tujih najpopularnejših modelih. Najboljše rezultate smo dobili z GaMS3-12B-Instruct, prispevek z natančno analizo in primerjavo rezultatov je sprejet na LLMs4SSH at LREC 2026 (Munda et al., v objavi).

S kombinacijo analize grafov imenskih entitet in kritične analize diskurza smo primerjali rabo in vrednotenje kolektivnih identitet v različnih temah in različnih časopisih, študija primera s podrobnim opisom metodologije in vizualizacijami je dostopna kot e-print na arXiv (Dobranić et al. 2026).