

Za: Društvo slovenskih književnih prevajalcev

Vprašanja glede zbiralne akcije besedil za veliki jezikovni model

Povezava na celoten dopis DSKP: <https://www.dskp-drustvo.si/wp-content/uploads/2025/03/Vprasanja-DSKP-v-zvezi-z-zbiralno-akcijo-besedil-PovejmoSi-final.pdf>

Pri književnih prevajalcih in izvirnih avtorjih gre za avtorska razmerja, ki jih ureja Zakon o avtorski in sorodnih pravicah. Besedila so avtorsko zaščitena, ob čemer tovrstno zbiranje ne spada pod izjemo »podatkovnega rudarjenja«. Kako v tem primeru ravnate z avtorskimi pravicami? Kako zaščitite dela pred nepooblaščenim rabo in širjenjem?

Projekt Povejmo izvajamo na Fakulteti za računalništvo in informatiko Univerze v Ljubljani skupaj s sedmimi partnerji. Cilj projekta je izdelava odprtega jezikovnega modela za slovenščino, ki je dostopen vsem na enak način pod enakimi pogoji. Hkrati v okviru Centra za jezikovne vire in tehnologije vzpostavljamo tudi tehnološko infrastrukturo, ki naj bi zagotovila, da bo v aplikacijah umetne inteligence lahko na nadzorovan način trajno podprta tudi slovenščina. V ta namen smo vzpostavili in zagotavljamo varno hrambo besedil v slovenskem jeziku, tudi avtorsko zaščitene, ter pravne podlage in tehnične pogoje za vključevanje teh besedil v učenje našega ter drugih odprto dostopnih velikih jezikovnih modelov za slovenščino.

Pri pravnih rešitvah sledimo trenutno veljavni slovenski in evropski zakonodaji glede upravljanja avtorskih pravic. Ob pomoči pravnih strokovnjakov smo tako pripravili Dogovor o uporabi avtorskih del v zvezi z zbiranjem besedil za namen gradnje odprtih jezikovnih modelov za slovenski jezik. Ta dogovor je dostopen na spletni strani povejmo.si in ga uporabljamo pri zbiranju besedil, ki so vključena v učenje slovenskega jezikovnega modela GaMS. Avtorji oziroma lastniki avtorskih pravic, ki svoja besedila prispevajo, se torej eksplicitno strinjajo z njihovo rabo za ta specifičen namen. V primerih, ko nam dostop omogočajo ustanove, ki gradiva le hranijo, niso pa lastniki avtorskih pravic, kot so na primer knjižnice, iščemo tudi drugačne pravno vzdržne rešitve.

Za nas je ključno, da vzpostavimo trajen in robusten sistem zbiranja in hranjenja slovenskih besedil v obliki, primerni za učenje jezikovnih modelov, ki mu lastniki avtorskih pravic, ki se s tako uporabo besedil strinjajo, lahko tudi zaupajo. Šele obstoj take infrastrukture lahko osmisli nadaljnje odločanje, katera avtorskopravno zaščitena gradiva bodo dejansko vključena v učenje bodočih jezikovnih modelov in katera ne. To bo tako ali drugače določala sedanja in bodoča zakonodaja ter tudi širši družbeni konsenz, tako v Sloveniji kot v EU.

Študijo glede izjeme podatkovnega rudarjenja (Copyright Law and Generative AI Training - Technological and Legal Foundations), ki jo navajate, je naročila in financirala evropska Pobuda za avtorske pravice. Predstavlja eno od strokovnih mnenj na področju, na katerem se mnenja sicer intenzivno krešejo. Glede uporabe avtorsko zaščitene gradiva bomo sledili veljavni zakonodaji in tako je zasnovan tudi trenutni Dogovor o uporabi avtorskih del v zvezi z zbiranjem besedil za namen gradnje odprtih jezikovnih modelov za slovenski jezik. Če se bo evropska ali slovenska zakonodaja tako ali drugače spremenila, tudi na podlagi te ali podobnih študij, bomo spremembe upoštevali.

Ali za uporabo besedil (npr. celotne knjige) predvidevate denarno nadomestilo? Ali bodo jezikovna orodja, ki bodo nastala na podlagi prejetih gradiv, dostopna vsem, brez omejitev? Lahko jamčite, da jih ne boste v prihodnosti tržili?

Če to želimo ali ne, trenutno poteka implicitno tekmovanje med jeziki in tudi tekma s časom, ko se odloča, do katere mere bo ta ali oni jezik podprt z umetno inteligenco. Pri tem je najpomembnejše vprašanje, do kakšne količine podatkov v določenem jeziku bodo imeli dostop veliki jezikovni modeli

pri učenju. Jeziki z večjim številom govorcev imajo pri tem povsem nenadomestljivo prednost pred jeziki, kot je slovenščina. Če zberemo vsa besedila, ki so bila kdajkoli v zgodovini napisana v slovenščini, bodo predstavljala le droben delček v primerjavi z angleškimi, kitajskimi, španskimi in drugimi besedili, na katerih so bili naučeni veliki jezikovni modeli. Dejansko gre pri tem vprašanju za kolektivno odločitev govorcev in govork slovenščine, tako v slovenski državni skupnosti, pa tudi v širšem okviru Evropske unije, do katere mere želimo, da bo umetna inteligenca znala slovensko oziroma bo poznala slovensko okolje in kulturo. Če si to želimo, moramo pravzaprav zbrati večino besedil v slovenščini ter na nadzorovan in konsenzualen način omogočiti njihovo vključitev v jezikovne modele.

Namen Centra za jezikovne vire in tehnologije je zagotavljati osnovno infrastrukturo za slovenski jezik, ki je odprta in globalno dostopna vsem pod enakimi pogoji, tako za nekomercialne kot za komercialne namene. Menimo, da odprt jezikovni model, ki ga gradimo v okviru projekta Povejmo, predstavlja nujen del take jezikovne oziroma jezikovnotehnološke infrastrukture. Predpostavljamo, da vprašanje denarnih nadomestil postane relevantno, ko se odprti modeli uporabljajo v komercialne namene in to vprašanje naj se razrešuje med akterji znotraj komercialne sfere. Menimo, da bi morala biti osnovna jezikovna infrastruktura dostopna vsem pod enakimi pogoji, podobno kot smo v ustavo zapisali pravico do pitne vode. Jezik kot tak ne bi smel biti tržno blago, za osnovno jezikovno infrastrukturo pa bi morala poskrbeti država.

Ali imate vpogled v sklenjene avtorske pogodbe med avtorji in založniki? Kako boste pri založbah preverjali, ali avtor dovoljuje rabo svojih besedil za učenje UI?

Besedilodajalci oziroma lastniki avtorskih pravic – bodisi založbe ali avtorji oz. avtorice – ob podpisu Dogovora o uporabi avtorskih del v zvezi z zbiranjem besedil za namen gradnje odprtih jezikovnih modelov za slovenski jezik jamčijo, da imajo vse pravice na avtorskih delih, tako da lahko izvajalcu projekta podelijo vsa potrebna dovoljenja. Tega posebej ne preverjamo.

Zanima nas, o kakšni transparentnosti je govor in kakšno metodologijo zagotavljanja transparentnosti imate. Prav tako omenjate »spoštovanje etičnih standardov«. Na kakšen način boste spoštovali etične standarde?

Kot rečeno, pri pravnih rešitvah sledimo veljavni slovenski in evropski zakonodaji glede upravljanja avtorskih pravic. Dogovor o uporabi avtorskih del v zvezi z zbiranjem besedil za namen gradnje odprtih jezikovnih modelov za slovenski jezik upošteva obstoječo veljavno zakonodajo. Avtorji oziroma lastniki avtorskih pravic, ki svoja besedila prispevajo, se strinjajo z njihovo rabo za ta specifičen namen. Vzpostavljena jezikovna infrastruktura z varno hrambo besedil ter upoštevanjem pravnih podlag za vključevanje besedil v učenje odprto dostopnih velikih jezikovnih modelov za slovenščino lastnike avtorskih pravic varuje pred možnimi zlorabami.

Avtorji doslej nismo bili povabljeni k besedi oz. nismo imeli priložnosti prispevati svojih pogledov o brezplačnem prispevanju avtorskih del za gradnjo velikih jezikovnih modelov in učenja UI.

Pri projektu smo vedno pripravljeni na dialog z vsemi deležniki, saj je le tako mogoče priti do družbenega konsenza glede obsega in načina podpore slovenščini v okviru umetne inteligence. Projekt Povejmo je bil decembra 2024 predstavljen tudi na Slovenskem knjižnem sejmu, ki ga organizira Zbornica knjižnih založnikov in knjigotržcev, in sicer na okrogli mizi z naslovom Ali bo umetna inteligenca pisala kot Ivan Cankar? Okroglo mizo je organizirala in neposredno prenašala Radiotelevizija Slovenije, posnetek si je mogoče ogledati na MMC. Zbornica knjižnih založnikov in knjigotržcev do sedaj ni izrazila želje po dodatnih predstavitev ali pogovorih.

Vabilo Društva slovenskih književnih prevajalcev s konkretnimi datumi za srečanje, v katerega sta bila vključena tudi predsednica Društva slovenskih književnih prevajalcev in predsednik Društva slovenskih pisateljev, smo prejeli 14. marca 2025. Zaradi odsotnosti udeležencev žal še vedno iščemo termin, ki bi ustrezal vsem. Vedno pa se skušamo čim hitreje odzvati na vsa vabila za predstavitev projekta ali odgovoriti na vprašanja glede dilem, tudi avtorskopравnih.