UDK 811.163.42(091):004
Pregledni rad
Rukopis primljen 2. VII. 2025.
Prihvaćen za tisak 30. IX. 2025.
https://doi.org/10.31724/rihjj.52.1.25

Marko Tadić
Filozofski fakultet Sveučilišta u Zagrebu
Ivana Lučića 3, HR-10000 Zagreb
https://orcid.org/0000-0001-6325-820X
marko.tadic@ffzg.hr

HR-CLARIN ISTRAŽIVAČKA INFRASTRUKTURA KAO PODLOGA ZA ISTRAŽIVANJE POVIJESTI HRVATSKOGA JEZIKA

Ovaj rad opisuje nastanak CLARIN ERIC-a i HR-CLARIN-a, izlaže njihovu misiju i glavne zadatke kao i ulogu repozitorijā objedinjenih u sveeuropsku federaciju repozitorija jezičnih podataka. Nadalje, u radu se razmatra uloga HR-CLARIN-a kao središnjega mjesta za pohranu, prikupljanje, razvoj i dijeljenje jezičnih resursa i alata kako za sinkronijsku, tako i za dijakronijsku analizu hrvatskoga jezika. Daje se i nekoliko prijedloga u kojem bi se smjeru trebali razvijati hrvatski jezični resursi za istraživanje povijesti hrvatskoga jezika.

1. Uvod

Digitalizacija donosi nove mogućnosti istraživanja jezika i drugih o jeziku ovisnih i s jezikom povezanih tema unutar humanističkih i društvenih znanosti. S njom su neposredno povezani i izazovi digitaliziranja jezičnih izvora, njihove pohrane, obradbe kao i izazovi njihova opisa i uporabe. Opisu starijih stanja hrvatskoga jezika svakako može pridonijeti razvoj istraživačkih infrastruktura za hrvatski jezik. HR-CLARIN kao dio CLARIN ERIC-a, europske istraživačke infrastrukture za humanističke i za društvene znanosti, središnje je mjesto prikupljanja, razvoja i dijeljenja jezičnih resursa i jezičnih alata, kako za potrebe sinkronijske analize, tako i za potrebe dijakronijske analize hrvatskoga.

Ovaj rad opisuje nastanak CLARIN ERIC-a i HR-CLARIN-a, izlaže njihovu misiju i glavne zadatke kao i ulogu repozitorijā objedinjenih u sveeuropsku federaciju repozitorija jezičnih podataka. Nakon uvoda u drugom dijelu rada opisuje se nastanak CLARIN-a prvotno kroz pripremni projekt CLARIN, pa preko uspostave CLARIN ERIC-a, a potom se opisuje i CLARIN-ov Virtualni jezični opservatorij kao i obitelji jezičnih resursa. U trećem dijelu rada objašnjava se nastanak i uloga HR-CLARIN-a, njegov konzorcij i uspostavljeni repozitorij, kao i centar znanja o hrvatskome jeziku Croatina. U četvrtom dijelu rada razmatra se kako HR-CLARIN može pomoći proučavanju hrvatskoga jezika uključujući i dijakronijsku perspektivu, a rad završava zaključnim napomenama i mogućim budućim smjerovima razvoja.

2. Europska istraživačka infrastruktura CLARIN ERIC

CLARIN je pokrata za *Common LAnguage Resources and Technology INfrastructure (for Humanities and Social Sciences)*, tj. Zajednički jezični resursi i tehnološka infrastruktura za humanističke i društvene znanosti (HiDZ).[1] Riječ je o europskoj istraživačkoj infrastrukturi uspostavljenoj u skladu s planom razvoja istraživačkih infrastruktura definiranim Europskim strateškim forumom za istraživačke infrastrukture (ESFRI),[2] a s pomoću pravnoga instrumenta Konzorcij europskih istraživačkih infrastruktura (ERIC).[3] Unutar ESFRI-ja, koji obuhvaća istraživačke infrastrukture sa svih znanstvenih područja, iz humanističkih znanosti postoje samo CLARIN i DARIAH (uz dvije u nastajanju: OPERAS i EHRI), a iz društvenih znanosti ESS, CESSDA i SHARE.[4] Republika Hrvatska sudjeluje kao punopravna članica u svim europskim istraživačkim infrastrukturama iz HiDZ-a.

[1] https://www.clarin.eu/

[2] https://www.esfri.eu/

[3] https://eur-lex.europa.eu/legal-content/HR/TXT/HTML/?uri=CELEX:32009R0723&qid=1751383979335

[4] Pobliže o svim tim istraživačkim infrastrukturama može se saznati na odgovarajućim mrežnim stranicama: https://www.dariah.eu/, https://operas-eu.org/projects/operas-plus/, https://www.ehri-project.eu/, https://www.europeansocialsurvey.org/, https://www.cessda.eu/, https://share-eric.eu/.

2.1. Pripremni projekt CLARIN

U načelu je svaka europska istraživačka infrastrukura započela pripremnim projektom koji je najčešće nosio ime buduće infrastrukture. Tako je bilo i s CLARIN-om, koji se našao već na prvome ESFRI-jevu planu istraživačkih infrastruktura 2006. i isprva je bio projektom u okviru Sedmoga okvirnoga programa EU-a koji je trajao od siječnja 2008. do lipnja 2011. Kako je već tada razvoj jezičnih tehnologija toliko uznapredovao da su mogle ponuditi potporu istraživačima iz drugih HiDZ, odlučeno je da se jezične tehnologije upotrijebe kao svojevrsno pomagalo, tj. istraživačka infrastruktura koja bi u HiDZ mogla odigrati onu metodološku ulogu kakvu u prirodnim znanostima ima matematika. Naime, kod velikoga je broja disciplina u HiDZ-u izravni objekt proučavanja tekst, tj. jezik nekoga teksta, a to uključuje sve filologije, lingvistiku, fonetiku itd., ili im je objekt proučavanja posredovan tekstom kao što je to slučaj u teoriji književnosti, etnologiji, povijesti itd. Razvijene jezične tehnologije za HiDZ u mnogome pomiču metodološku strogost njihovih istraživačkih postupaka i omogućuju uvođenje paradigme e-znanosti, tj. znanosti koja će se baviti i računalno predstavljenim objektima i modelima funkcioniranja raznih sustava opstojaka u HiDZ-u.

Misija projekta CLARIN bila je stvoriti istraživačku infrastrukturu koja omogućuje uporabu jezičnih resursa (JR) i jezičnih tehnologija (JT) istraživačima ih svih znanosti, a osobito HiDZ-a. Načelno se to postiglo dosezanjem nekoliko ključnih ciljeva:

- ujedinjavanjem postojećih digitalnih repozitorija jezičnih podataka u europsku federaciju certificiranih digitalnih repozitorija
- osiguravanjem jedinstvenoga i jednoobraznoga pristupa svim repozitorijima
- formiranjem jezičnotehnoloških alata kao *web-servisa* za pretragu, rukovanje, obogaćivanje i korištenje podataka iz tih repozitorija
- ravnopravnim uključivanjem svih jezika
- omogućivanjem sudjelovanja svih država-članica i država-pristupnica za EU.

Projekt CLARIN okupio je 33 partnera iz 22 države i u svakoj od njih je uspio uključiti najznačajnije istraživačke ustanove s područja računalne lingvistike,

obrade prirodnoga jezika i jezičnih tehnologija, a Hrvatsku je predstavljao Filozofski fakultet Sveučilišta u Zagrebu. Namjena projekta CLARIN bila je planiranje i izgradnja prototipa istraživačke infrastrukture, tj. CLARIN-ove mreže repozitorija jezičnih podatka. Okosnicu te mreže čine tzv. B-centri, tj. pojedinačni repozitoriji koji pohranjuju jezične podatke i njihove metapodatke, a metapodatci se agregiraju i propagiraju kroz cijelu mrežu i dostupni su za pretragu svim korisnicima. Kako bi postao B-centar, svaki repozitorij mora proći certifikacijski postupak i osigurati trajnu dostupnost podataka u repozitoriju. Za razliku od B-centara, C-centri pohranjuju samo metapodatke i osiguravaju njihovu trajnu dostupnost.

2.2. CLARIN ERIC

Nakon završetka projekta CLARIN, daljnja se aktivnost konzorcija usredotočila na osnivanje CLARIN ERIC-a koji je osnovan u Haagu 18. travnja 2012. godine. Članice CLARIN ERIC-a mogu biti države-članice, države-pristupnice EU-u, ili nadnacionalne organizacije, pa su tako osnivačice CLARIN ERIC-a toga dana postale Austrija, Bugarska, Češka, Danska, Estonija, Nizozemska, Nizozemska jezična unija, Njemačka i Poljska.

Svaka je punopravna članica CLARIN ERIC-a preuzela obvezu financiranja nacionalnoga konzorcija CLARIN-a, izgradnju infrastrukture na nacionalnoj razini i povezivanje te infrastrukture s CLARIN ERIC-om na razini EU-a.

Istraživačka infrastruktura CLARIN ERIC osigurava:

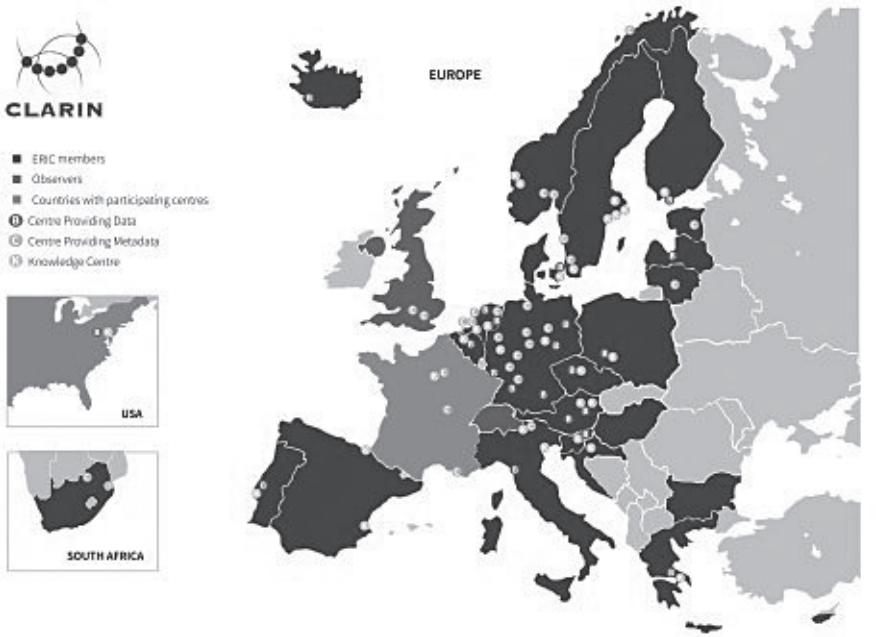
- trajni pristup skupovima podataka za znanstvenike u HiDZ-u
- digitalne jezične podatke u pisanom, govorenom ili multimodalnom zapisu
- alate za otkrivanje, analizu i kombiniranje podataka bez obzira gdje su pohranjeni
- okružje jedinstvene prijave (npr. aai@edu.hr za istraživače iz Hrvatske)

i zapravo zaokružuje ekosustav za razmjenu znanja koji je vertikalno uključen u EOSC-u (*European Open Science Cloud*)[5] i stoga slijedi FAIR (*Findable, Accessible, Interoperable, Re-usable*) načela pristupa podatcima (Eskevich i

[5] https://open-science-cloud.ec.europa.eu

dr. 2020: 29). Svaki jezični resurs u CLARIN ERIC-ovoj federaciji repozitorija dobiva jedinstven i trajan identifikator (PID) i može se navoditi kao podatkovna referencija u znanstvenim radovima. Time se konačno omogućuje i vrjednovanje često zanemarenih mnogih istraživačkih napora pri prikupljanju, ustavljivanju i priređivanju jezičnih podataka, a koji su napori u načelu bili nepriznavani kao značajan rad u struci.

CLARIN ERIC danas broji 26 članica, 22 B-centra, 32 C-centra i 37 K-centara.[6] K-centri (*Knowledge centres*) centri su izvrsnosti i specijaliziranih znanja o pojedinim temama ili jezicima. U načelu svaki nacionalni konzorcij otvara K-centar za nacionalni jezik, pa je tako otvoren i K-centar za hrvatski jezik Croatina (v. više u dijelu 3.3.).



Slika 1: Prikaz stanja europske istraživačke infrastrukture CLARIN ERIC prema zemljama i vrstama centara krajem 2024. (preuzeto s https://www.clarin.eu/content/national-consortia)

[6] https://www.clarin.eu/content/national-consortia

Ako se za beskonačno veliki makrokozmos koriste teleskopi i za beskonačno malen mikrokozmos koriste mikroskopi, na CLARIN ERIC bi se moglo gledati kao na makroskop s pomoću kojega se mogu promatrati ponekad beskonačno složene pojave poput ljudskoga društva, odnosa u njemu, odnosa između pojedinaca i društvenih skupina, odnosa između pojedinaca i pojedinaca i sve to posredovano jezikom.

2.3. CLARIN-ov Virtualni jezični opservatorij

Takav markoskopski pogled na jezične podatke i sve odnose koje oni kodiraju omogućuje upravo CLARIN ERIC-ov sustav za pretragu pod nazivom Virtualni jezični opservatorij (VLO).[7] On omogućuje vrlo preciznu pretragu ukupnoga skupa metapodataka koju odlikuje tzv. fasetna pretraga u kojoj se odmah pojavljuje mogući skup odgovora na tražen upit prema kategorijama s brojem odgovora, pa je upit moguće usmjereno sužavati. VLO za svaki pronađeni jezični resurs daje poveznicu na njegovu mrežnu stranicu, mogućnost njegova preuzimanja, uvid u licencije za njegovu uporabu i u njegove tehničke značajke, pregled alata koji se mogu koristiti za obradu toga jezičnoga resursa kao i uputu kako ga citirati. Takvo izrazito fleksibilno sučelje za pretragu čitave federacije repozitorija omogućuje VLO-u visoku učinkovitost.

2.4. CLARIN-ove obitelji resursa

CLARIN ERIC je uveo na makrorazini još jedan oblik usustavljivanja jezičnih resursa, a to su tzv. obitelji jezičnih resursa (Fišer i dr. 2018). Obitelji CLARIN-ovih jezičnih resursa daju jednostavan pregled JR-a prema vrsti podataka dostupnih u njima u CLARIN-ovoj infrastrukturi za istraživače iz digitalnih HiDZ.

Svi su jezični resursi smješteni u konačan skup kategorija i prema njima se može lakše pronaći onaj resurs za onaj jezik i ono razdoblje ili onu domenu koja je istraživačima potrebna. Na najvišoj razini ta se klasifikacija jezičnih resursa sastoji se od tri kategorije: korpusi, leksički resursi i alati. Na nižoj razini među

[7] https://vlo.clarin.eu/

korpusima nalaze se npr. korpusi znanstvenih tekstova, korpusi starih tekstova, pravni korpusi, beletristički korpusi, novinski korpusi, korpus parlamentarnih rasprava, usporedni korpusi, referentni korpusi itd. Među leksičkim resursima nalaze se npr. rječnici, glosariji, pojmovnici, veliki jezični modeli (VJM-i), popisi riječi itd. Među jezičnim alatima nalaze se npr. alati za pretragu korpusa, za prepoznavanje imena, za označavanje vrsta riječi i gramatičkih kategorija, za lematizaciju, za parsanje, za analizu mnijenja i stavova itd.

3. HR-CLARIN

Predstavnik Hrvatske bio je pozvan sudjelovati u osnivanju CLARIN-a od samoga početka, tj. od pripremnoga neformalnoga sastanka u Parizu 6. veljače 2006. i koordinacijskoga sastanka u Budimpešti 30. i 31. listopada 2006. To je posljedično dovelo do sudjelovanja Filozofskoga fakulteta Sveučilišta u Zagrebu kao partnera u pripremnom projektu CLARIN.

Premda je Hrvatska u prosincu 2010. potpisala memorandum o razumijevanju, na žalost, nije se našla među državama osnivačicama CLARIN ERIC-a jer je nakon promjene administracije i političkoga zaokreta u siječnju 2012. novi ministar znanosti zaustavio sve aktivnosti oko pristupanja Hrvatske u ERIC-e, a to je posljedično dovelo do zastoja u pristupanju Hrvatske i CLARIN ERIC-u. To je spriječilo Hrvatsku da u potpunosti sudjeluje u CLARIN ERIC-u do 2018. godine kada je, nakon nekoliko iteracija pregovora s četirima različitim ministrima tijekom sedam godina, konačno donesena odluka o sudjelovanju Hrvatske u CLARIN ERIC-u kao punopravne članice. Uvjereni smo da nismo bili jedina zemlja sa sličnim tretmanom od strane njezinih odgovarajućih ministarstava, agencija, vijeća ili zaklada, ali vjerujemo da se hrvatski put u CLARIN svakako može smatrati jednim od najsloženijih. Međutim, ova odluka iz 2018. bila je više deklarativne naravi jer je hrvatski CLARIN morao čekati još dvije godine da se osigura financiranje nacionalnoga konzorcija koje bi omogućilo provođenje svih potrebnih radnji za pokretanje nacionalnoga konzorcija HR-CLARIN[8] i koordinaciju svojih aktivnosti.

[8] https://clarin.hr

3.1. Konzorcij HR-CLARIN

Konzorcij HR-CLARIN sastoji se od osam članova. Uz Filozofski fakultet Sveučilišta u Zagrebu (FFZG), koji je koordinator, to su još: Sveučilišni računski centar (SRCE), Fakultet elektrotehnike i računarstva (FER), Institut za hrvatski jezik (IHJ), Nacionalna i sveučilišna knjižnica (NSK), Hrvatska akademija znanosti i umjetnosti – Knjižnica (HAZU) i Hrvatsko društvo za jezične tehnologije (HDJT). Konzorcij je otvoren za druge ustanove u Hrvatskoj koje proizvode ili pohranjuju jezične resurse, a među prvima koji su iskazali zanimanje za uključivanje su Edukacijsko-rehabilitacijski fakultet Sveučilišta u Zagrebu i Filozofski fakultet Sveučilišta u Rijeci.

3.2. Repozitorij HR-CLARIN-a

Uzimajući u obzir određene specifičnosti metapodatkovnoga opisa koji je standardiziran unutar CLARIN ERIC-a, razvojni je tim LINDAT-CLARIAH/CZ centra s Karlovoga sveučilišta u Pragu razvio inačicu popularnoga repozitorijskoga sustava DSpace specifičnu za CLARIN (Hajič i dr. 2022). Ta je inačica vrlo dobro prihvaćena unutar CLARIN ERIC-ove zajednice, pa se za nju se odlučio i konzorcij HR-CLARIN.

Nakon što su zadovoljeni svi tehnički preduvjeti, uspostavljen je repozitorij HR-CLARIN-a,[9] čime je i hrvatskim znanstvenicima otvoren pristup pohrani i dijeljenju jezičnih resursa, ponajprije na hrvatskome, ali i na drugim jezicima. Korisnici u repozitorij mogu pohranjivati svoje jezične resurse, pri čemu svaki jezični resurs dobiva jedinstveni trajni identifikator (PID) i preporučuje ga se uključiti u citiranje korištenih jezičnih resursa. Budući da je citiranje podatkovnih izvora još uvijek neriješeno pitanje unutar hrvatske znanstvene zajednice, CLARIN nudi i preporuke najbolje prakse citiranja koristeći trajne identifikatore kako bi, s jedne stane rad autora jezičnih podataka bio adekvatno vrjednovan, a s druge strane, kako bi se korišteni skup podataka jednoznačno odredio u svrhu osiguravanja ponovljivosti istraživanja (Farkaš i dr. 2024).

[9] https://repository.clarin.hr

Hrvatski jezični resursi pohranjeni u repozitoriju HR-CLARIN-a izravno su uključeni u europsku federaciju repozitorija CLARIN ERIC i vidljivi su i, ovisno o pridruženoj licenciji, dostupni svim autoriziranim istraživačima i/ili široj zainteresiranoj javnosti.

Prije uspostave samoga repozitorija i s ciljem izrade kataloga jezičnih resursa i alata koji će se smjestiti u repozitorij HR-CLARIN-a, konzorcij je pokrenuo kampanju prikupljanja (meta)podataka o jezičnim resursima i alatima izrađenima unutar hrvatske znanstvene zajednice. Kampanja prikupljanja podataka zainteresiranih za uključivanje svojih resursa u repozitorij HR-CLARIN-a trajno je otvorena i svi koji se žele uključiti pozvani su ispuniti anketni upitnik.[10]

U ovom je trenutku repozitorij HR-CLARIN-a u statusu C-centra, ali je pokrenut postupak certifikacije za status B-centra.

3.3. Centar znanja Croatina

Croatina[11] ili CLARIN-ov centar znanja (tzv. *K-centre*) za hrvatski jezik osnovan je 2024., a u njemu sudjeluju dvije ustanove koje su obje članice nacionalnog konzorcija HR-CLARIN: Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu (FFZG) i Institut za hrvatski jezik (IHJ). Croatina pruža relevantna znanja o hrvatskom jeziku i promiče korištenje jezičnih tehnologija za hrvatski jezik, korisnicima daje potporu putem službe za korisnike (*helpdesk*) pružajući relevantne informacije o temama vezanim uz hrvatski jezik, daje savjete korisnicima o izgradnji i pohrani vlastitih jezičnih resursa (Farkaš i dr. 2024). Korisnicima Croatine nudi se pomoć s pristupom podatcima i alatima te s uporabom jezičnih resursa, a također se daju preporuke o širokom spektru tema vezanih uz jezične tehnologije, ponajprije za hrvatski jezik. Croatina pokriva ponajprije lingvističke teme i obradu prirodnoga jezika. Nudi se specifična ekspertiza iz različitih vrsta podataka o hrvatskom jeziku, ali i opće teme kao što su upravljanje podatcima i metapodatcima, jezičnotehnološke platforme, strojno potpomognuto učenje hrvatskoga kao stranoga jezika itd. S obzirom na to da Croatina pruža kontinuiranu podršku korisnicima, svi se oni mogu obratiti

[10] Obrazac za prijavu jezičnih resursa i tehnologija stvorenih unutar hrvatske znanstvene zajednice dostupan je na poveznici https://limesurvey.srce.hr/549263?lang=hr.

[11] https://www.clarin.hr/hr/centri/croatina/

za sva pitanja vezana za teme koje centar znanja Croatina pokriva putem službe za korisnike ili adrese e-pošte.[12]

4. Kako HR-CLARIN može pomoći u proučavanju hrvatskoga jezika?

HR-CLARIN ponajprije može pomoći u proučavanju hrvatskoga jezika svojom ulogom središnjega repozitorija jezičnih podataka koji je ujedno uključen u europsku istraživačku infrastrukturu i time je omogućena vidljivost hrvatskih jezičnih resursa svim istraživačima s pristupom CLARIN ERIC-u. Istodobno, hrvatski istraživači imaju pristup čitavoj europskoj federaciji repozitorija jezičnih podataka i do sada nezamislivim mogućnostima njihove pretrage i uporabe u svojim istraživanjima.

Repozitorij HR-CLARIN-a daje mogućnost besplatne trajne pohrane jezičnih resursa kao što su korpusi, rječnici i veliki jezični modeli, ali i jezičnih alata kao što su sustavi za lematizaciju, PoS/MSD-označavanje, sustava za parsanje itd. Već se sada u repozitoriju HR-CLARIN-a nalazi prvi hrvatski veliki jezični model HR-GPT Beta, ali i doobučeni višejezični modeli HR-Gemma Beta i HR-Pythia Beta (Thakkar i dr. 2024).

Dok se situacija s jezičnim resursima u sinkronijskom presjeku čini relativno dobro razvijenom (Tadić 2022, Tadić 2023), u dijakronijskom presjeku, tj. u jezičnim resursima za starija stanja hrvatskoga jezika dosta je nedostataka. Nedovoljno je digitaliziranih tekstova i korpusa premda su se u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu još sedamdesetih i osamdesetih godina korpusno obrađivali stari pisci hrvatski (Tadić 1997). Digitaliziran je zadovoljavajući broj starih hrvatskih rječnika dostupnih na Portalu hrvatske rječničke baštine[13] (Bago i Boras 2012) kao i rječnika nastalih i na dijakronijskoj građi u Institutu za hrvatski jezik,[14] ali ne jednoobrazno. Svakako bi u okviru konzorcija HR-CLARIN-a trebalo koordinirati popunjavanje

[12] Mrežna poveznica službe za korisnike je https://www.clarin.hr/hr/centri/croatina/helpdesk/, a adresa e-pošte helpdesk@clarin.hr.

[13] http://crodip.ffzg.hr/

[14] v. npr. https://kajkavski.hr/

zapisa u repozitoriju HR-CLARIN-a s opsežnim digitalizacijskim kampanjama koje provode NSK i Knjižnica HAZU-a kako bi se rezultati tih kampanja mogli pronaći u čitavom CLARIN ERIC-u. U tim kampanjama obuhvaćene su velike količine publikacija i dokumenata zapisanih starijim stanjima hrvatskoga jezika i s različitim slovopisnim i pravopisnim rješenjima. Za proučavanje takvih dokumenata i publikacija valjalo bi razviti sustav za automatsku transkripciju hrvatskih tekstova sa starijih slovopisnih i pravopisnih rješenja na današnji standardnojezični pravopisni uzus. Jedna od metoda zacijelo bi mogla biti izrada dijakronijskoga velikoga jezičnoga modela za hrvatski jezik, a takvih modela za starija stanja hrvatskoga jezika uopće još nemamo. Tu bi od pomoći mogao biti novoosnovani CLARIN-ov K-centar LLMs4SSH[15] koji za cilj ima pružati pomoć i ekspertizu za uporabu VJM-a u HiDZ-u. Trebalo bi provesti niz eksperimenata u kojima bi se utvrdilo je li možda brže i jednostavnije doći do takva sustava dodatnom obukom nekoga od postojećih sinkronijskih VJM-a za hrvatski jezik i time ga osposobiti za primjenu i na starijim hrvatskim tekstovima.

5. Zaključci i buduća istraživanja

Predstavljena je europska istraživačka infrastruktura CLARIN ERIC i hrvatsko sudjelovanje u njoj – HR-CLARIN. Hrvatski čvor u toj europskoj istraživačkoj infrastrukturi predstavlja središnje mjesto za pohranu, dijeljenje i uporabu hrvatskih jezičnih resursa kako suvremenoga hrvatskoga, tako i jezičnih resursa starijih stanja hrvatskoga jezika. U tome se može pronaći njegova najveća vrijednost.

U planovima za buduća istraživanja i daljnji razvoj HR-CLARIN-a nalaze se ponajprije namjere za osnivanje dva nova K-centra: za srednjovjekovni i novovjekovni latinizam i za glagoljicu. Smatramo kako je upravo Hrvatska ono mjesto u Europi koje može pružiti vrijedne znanstvene doprinose u oba ova područja, pa će se inicijative za osnivanje tih K-centara dodatno pospješivati. Uz to unutar HR-CLARIN-a provodi se izradba novih jezičnih resursa (npr. Hrvatski znanstveni korpus) ili novih inačica postojećih resursa (npr. Hrvatski nacionalni korpus v4) koji bi u dogledno vrijeme morali zauzeti svoje zasluženo mjesto.

[15] https://llms4ssh.clarin-pl.eu/

Kako god bilo, HR-CLARIN je potpuno ekipiran i spreman udomiti rezultate rada svih hrvatskih jezikoslovaca koji proizvode bilo kakve digitalne jezične podatke i resurse i žele ih učiniti dostupnima u cijeloj europskoj istraživačkoj infrastrukturi.

Zahvale

Konzorcij HR-CLARIN podupire Ministarstvo znanosti, obrazovanja i mladih Republike Hrvatske.

Literatura

Bago, Petra; Boras, Damir. 2012. *A framework for consolidating most important digitized Croatian dictionaries from 1595 to 1945*. Prezentacija sa skupa *17. International Information Technology and Journalism Conference*.

Eskevich, Maria i dr. 2020. CLARIN: Distributed Language Resources and Technology in a European Infrastructure. *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)*. Ur. Rehm, Georg i dr. ELRA. Marseille. 28–34.

Farkaš, Daša; Štefanec, Vanja; Tadić, Marko. 2025. HR-CLARIN istraživačka infrastruktura i centar znanja za hrvatski jezik Croatina. *Jezik u doba globalizacije. Zbornik radova godišnje konferencije Hrvatskoga društva za primijenjenu lingvistiku*. Ur. Pon, Leonard; Šarić Šokčević, Ivana. HDPL. Osijek. 156–169.

Fišer, Darja; Lenardič, Jakob; Erjavec, Tomaž. 2018. CLARIN's Key Resource Families. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC2018)*. Ur. Calzolari, Nicoletta i dr. ELRA. Miyazaki. 1320–1325.

Hajič, Jan i dr. 2022. LINDAT/CLARIAH-CZ: Where we are and where we go. *CLARIN – The* Infrastructure for Language Resources. Ur. Fišer, Darja; Witt, Andreas. De Gruyter. Berlin – Boston. 61–82. https://doi.org/10.1515/9783110767377-003

Tadić, Marko. 1997. Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. *Suvremena lingvistika* 23 (43-44). 387–394.

Tadić, Marko. 2022. Report on the Croatian Language. *European Language Equality.* https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_7__Language_Report_Croatian_.pdf (pristupljeno 28. siječnja 2026.)

Tadić, Marko. 2023. Language Report Croatian. European Language Equality. Ur. Rehm, Georg; Way, Andy. Springer International Publishing. Cham. 111–114. http://dx.doi.org/10.1007/978-3-031-28819-7_9

Thakkar, Gaurish i dr. 2024. Building a Large Language Model for Moderately Resourced Language: A Case of Croatian. *Proceedings of the 35th Central European Conference on Information and Intelligent Systems (CECIS 2024)*. Ur. Vrček, Neven; Korent, Dina. University of Zagreb Faculty of Organization and Informatics. Varaždin. 225–232.

HR-CLARIN Research Infrastructure as a Basis for Researching the History of Croatian Language

Abstract

This paper describes the creation of CLARIN ERIC and HR-CLARIN, outlines their mission and main tasks as well as the role of repositories of language data integrated into the pan-European Federation of language repositories. In addition, the role of HR-CLARIN as a central place for storing, collecting, developing and sharing linguistic resources and tools for both synchronous and diachronous analysis of the Croatian language is considered in the paper. Several suggestions are also made as to the direction in which Croatian language resources for research into the history of the Croatian language should be developed.

Ključne riječi: istraživačke infrastrukture, CLARIN ERIC, HR-CLARIN, centar znanja Croatina, repozitorij

Keywords: research infrastructures, CLARIN ERIC, HR-CLARIN, knowledge centre Croatina, repository