FILOLOGIJA 84, Zagreb 2025.

UDK 811.163.42'33+004.85
DOI https://dx.doi.org/10.21857/y26kecd1n9
Izvorni znanstveni rad
Rukopis primljen 26. V. 2025.
Prihvaćen za tisak 20. X. 2025.

Marko Tadić
Odsjek za lingvistiku
Filozofski fakultet u Zagrebu
Ivana Lučića 3, HR-10000 Zagreb
marko.tadic@ffzg.hr

ZAŠTO SU JEZIČNE TEHNOLOGIJE BITNE ZA BUDUĆNOST HRVATSKOGA JEZIKA?

U radu se daje definicija, sastav i pregled jezičnih tehnologija kao i rezultati dvaju velikih kampanja ocjene stanja razvoja jezičnih tehnologija za više desetaka europskih jezika u kojima je sudjelovao i hrvatski. Dok je u kampanji META-NET-a u 2011. hrvatski smješten u skupinu jezika s nedostatno razvijenim jezičnim tehnologijama, u kampanji *European Language Equality* 2022. našao se u skupini od dvadesetak jezika s djelomično razvijenim jezičnim tehnologijama i pokazao se relativan napredak. Međutim, u razvoju jezičnih tehnologija upravo su veliki jezični modeli uveli promjenu paradigme i pokazali kako će se veliki broj do sada razvijenih jezičnih alata morati iznova proizvesti s novom metodologijom u pozadini tj. onom koja koristi velike jezične modele. Dalje se u radu daje pregled osnovnih vrsta velikih jezičnih modela i objašnjava se razlika između umjetne inteligencije i velikih jezičnih modela. Rad se zaključuje ukazivanjem na potrebu trajnoga razvoja jezičnih tehnologija za hrvatski jezik i to upravo na temelju stalnoga razvoja novih velikih jezičnih modela za hrvatski jezik u skladu sa svakom njihovom novom arhitekturom koja se bude pojavila. Za to je potrebno osigurati do sada neviđene količine tektovnih podataka na hrvatskom jeziku. U protivnom, hrvatskom se jeziku može dogoditi "digitalna nepismenost" i ostanak onkraj digitalne razdjelnice.

1. Uvod

Od javne objave prvoga velikoga jezičnoga modela slobodno dostupnoga općoj javnosti – ChatGPT-ja – u studenome 2022., jezične su tehnologije postale svakodnevica ne samo specijalistima i znalcima iz pod-

ručja računalnoga jezikoslovlja i informacijskih znanosti, već su na velika vrata ušli u profesionalne i privatne živote mnogih korisnika. Upravo je ChatGPT učinio uslugu svim računalnim jezikoslovcima u svijetu jer sad svi znaju čime se računalno jezikoslovlje i računalna obradba prirodnoga jezika (*Natural Language Processing, NLP*) zapravo bave i do kakvih se rezultata došlo. Kako bi se razumjela ta promjena u svojoj punini, na početku ovoga rada valjalo bi je smjestiti u slijed prethodnih istraživanja i tako odrediti što su to jezične tehnologije (JT), od čega se sastoje i kako nam mogu poslužiti.

U Hrvatskome općem leksikonu **tehnologija** je definirana kao »skup metoda i postupaka za preradbu sirovina u proizvode« (*Hrvatski opći leksikon*, 2012). Takva je definicija izravno primjenljiva na područje kemijske tehnologije gdje se sasvim jasno razaznaje što su sirovine (razne tvari, npr. SO_2 i H_2O), što su postupci (analiza ili sinteza tih tvari u nove tvari, npr. otapanje SO_2 u H_2O uz katalizator V_2O_5), a što proizvodi (nove se tvari mogu prodavati na tržištu, npr. H_2SO_4 tj. sumporna kiselina). Slično je razvidna primjenljivost ove definicije i u drugim tehnologijama, npr. nuklearnoj, informacijskoj, prometnoj, itd.

Međutim, kad je riječ o JT-ama, valja napomenuti kako one, poput današnjih komunikacijskih tehnologija (KT), u cijelosti ovise o informacijskim tehnologijama (IT). Dakle, bez IT-a nema ni JT-a. U tom smislu valja promatrati i sirovinu za JT-e, a to su **podatci o jeziku**, dakle, jezik pohranjen u digitalnom obliku ili u obliku e-teksta. Za proizvode JT-a ipak nema tako jednostavne i razvidne definicije, već bi ih se moglo opisati kao **sustave koji korisniku omogućuju jednostavn(ij)u uporabu prirodnoga jezika u digitalnom okružju** (Tadić 2003:13). Pri tome se digitalno okružje ne sastoji samo od uporabe teksta u računalima, već je ono znatno zastupljenije sveprisutnošću prijenosnih uređaja tj. mobitela koji su ništa drugo doli visokospecijalizirana računala za komunikaciju. Ti sustavi kao proizvodi JT-a zapravo zamjenjuju dio ljudskoga rada strojnim radom u raznim komunikacijskih situacijama. Primjer za to su npr. sustavi za strojno prevođenje gdje se rad čovjeka-prevoditelja zamijenio (više ili manje kvalitetno) strojnim "radom".

Iz tako definiranih JT-a proizlazi kako su podatci o jeziku tj. mnogobrojni prikupljeni e-tekstovi na nekom jeziku temeljni za razvijanje JT-a za taj jezik i bez dovoljne količine tako prikupljenih e-tekstova ne treba se niti upuštati u razvoj JT-a za taj jezik. Neka jezična zajednica mora doseći stupanj svijesti o samome svom jeziku, svojim potrebama za njim u suvremenim komunikacijskim kanalima i njegovoj uporabivosti u njima kako bi mogla početi ulagati u razvoj JT-a za svoj jezik.

Gore postavljena polazna i krajnja točka procesa prerade sirovina u proizvode kod JT-a zahtijeva daljnje sažeto razjašnjenje tj. podjelu na potpodručja. U Tadić (2003:27–28) daje se načelna podjela JT-a:

1. **jezičnih resursa:** računalni korpusi i digitalni rječnici;
2. **jezični alati** s pomoću kojih se jezični podatci (u korpusima i rječnicima) prerađuju u proizvode JT-a: razvijaju se i primijenjuju na svim jezičnim razinama tj. morfološkoj (opojavničenje, lematizacija, označavanje vrsta riječi i ostalih gramatičkih kategorija); sintaktičkoj (parsanje tj. sintaktička analiza rečenica, prepoznavanje imena i posebnih izraza u rečenicama); semantičkoj (prepoznavanje značenja riječi, prepoznavanje semantičkih uloga, prepoznavanje stavova govornika/pisca); pragmatičkoj (analiza i razrješenje deiktika, prepoznavanje govornih činova, komunikacijskih maksima i implikature), itd.;
3. (komercijalni) **proizvodi i** (mrežno dostupne) **usluge:** kao najpopularniji mogu se navesti provjernici pravopisa, gramatike i stila (*checkers*); digitalni rječnici, leksičke i terminološke baze; sustavi za automatsko indeksiranje deskriptorima ili sažimanje dokumenata; sustavi za obradu govora (automatsko prepoznavanje (*automatic speech recognition, ASR*) ili sinteza (*text to speech, TTS*)); strojno (potpomognuto) prevođenje (*machine translation, MT* i *machine assisted translation, MAT* / *computer assisted translation, CAT*); strojno potpomognuto učenje jezika (*computer assisted language learning, CALL*); dijaloški sustavi (*question/answering, Q/A*), itd.

Svi su nam ti sustavi danas dostupni samostalno ili integrirani u razne uređaje ili računalne pakete, a neke od njih koristimo i ne znajući kako pripadaju području jezičnih tehnologija, poput npr. sustava T9 koji nam omogućuje automatsku dopunu riječi pri unosu teksta kod pisanja SMS-ova ili poruka u raznim komunikacijskim aplikacijama na mobilnim uređajima.

U ovom će radu dalje biti riječi o stanju razvoja JT-a za hrvatski jezik u 2011. i 2021., o značajnoj promjeni paradigme u računalnoj obradi prirodnoga jezika uvođenjem velikih jezičnih modela (VJM), a zaključit će ga se ukazivanjem na to da razvoj JT-a za hrvatski jezik mora biti trajna zadaća.

2. Stanje razvoja jezičnih tehnologija za hrvatski jezik 2011.

Premda su se prvi koraci u razvoju onoga što danas nazivamo JT, u Hrvatskoj pojavili vrlo rano – već je 1959. u organizaciji profesora Bulcsúa Lászlóa i Svetozara Petrovića u Zagrebu organizirana prva radionica o strojnom prevođenju (László–Petrović 1959; Skansi–Mršić–Skelac 2020) – stanje razvijenosti JT-a za hrvatski jezik u 2011. još uvijek nije bilo zadovoljavajuće. Naime, u europskoj mreži izvrsnosti META-NET[1], koja se sastojala od četiriju EU-projekata, poduprtoj unutar Sedmoga okvirnoga programa EU-a, napravljene su tijekom 2011. snimke stanja razvoja JT-a za tridesetak europskih jezika, a hrvatski je smješten u skupinu s dvadesetak drugih jezika s izrazito nisko razvijenim JT-ama. Razlozi takvoj nedostatnoj razvijenosti JT-a za hrvatski bili su višestruki, a ponajvažniji su bili s jedne strane nedovoljna količina prikupljenih podataka o jeziku, a s druge strane relativno oskudna financijska potpora i nedovoljan broj istraživača. Ne treba smetnuti s uma da je tome pridonijela činjenica kako se nerijetko na istraživanja samostalnoga hrvatskoga jezika u SFRJ gledalo ne blagonaklono, nego kao sredstvo potkopavanja totalitarne komunističke diktature, pa su rijetki priručnici sa samostalnim imenom jezika (npr. znameniti Londonac) bili cenzurirani i uništavani, a autori na mnoge načine ograničavani i nerijetko kažnjavani. Situacija se promijenila stjecanjem samostalosti Republike Hrvatske, ali je trebalo još pričekati do druge polovice devedesetih godina 20. stoljeća jer su svi znanstveni odnosi EU-a s RH bili zamrznuti zbog osloboditeljskoga Domovinskoga rata. Pravi su se pomaci počeli događati tek krajem staroga i početkom novoga tisućljeća, a moglo bi se tvrditi kako je prekretnicu 1998. napravio Hrvatski nacionalni korpus[2] (Tadić 2009) kao treći veliki reprezentativni korpus za cjelinu nekoga jezika u povijesti korpusne lingvistike (prva dva su Britanski nacionalni korpus 1990. i Češki nacionalni korpus 1996.), ali zato prvi koji je bio slobodno mrežno pretraživ već godine 1999.

Kao jedan od rezultata META-NET-a, godine 2012. objavljen je niz tzv. Bijelih knjiga META-NET-a (Rehm–Uszkoreit 2012) u kojima su predstavljeni pregledi stupnja razvoja JT-a za europske jezike. Napravljena je snimka stanja za 30 jezika, a hrvatski je jezik smješten u skupinu jezika s izrazito nisko razvijenim JT-ama. Podrobnije o tome v. u Tadić–Brozović-Rončević–Kapetanović 2012, a ovdje ćemo samo navesti kako su svi relevantni podatci o stanju JT-a za hrvatski (uz bugarski, madžarski, poljski, slovački i

[1] *https://meta-net.eu/*.
[2] *https:// hnk.ffzg.hr/*.

srpski) prikupljeni unutar projekta *Central and Eastern European Resources (CESAR)*[3] kao jednoga od četiriju projekata unutar META-NET-a.

Za hrvatski su jezik tada bile na raspolaganju sljedeće JT-e:

- jezični resursi
  - korpusi
    - ◇ jednojezični: Hrvatski nacionalni korpus v2[4], Hrvatski jezični korpus (Riznica) Instituta za hrvatski jezik i jezikoslovlje[5], Hrvatski mrežni korpus v1[6] (hrWaC), a ostali su korpusi bili manji (npr. Hrvatska ovisnosna banka stabala, HOBS[7]) ili stariji (npr. Jednomilijunski korpus hrvatskoga književnoga jezika, tzv. Mogušev korpus, Katančićev prijevod Biblije, itd.)
    - ◇ višejezični: Hrvatsko-engleski usporedni korpus, Hrvatsko-engleski mrežni korpus[8] (hrenWaC), SETimes korpus[9]
  - rječnici i leksičke baze: Hrvatski morfološki leksikon[10], Hrvatski jezični portal[11], Portal hrvatske rječničke baštine[12] (Boras 2005), tezaurus EUROVOC[13], Četverojezični rječnik prava EU-a[14], Hrvatski WordNet (CroWN)[15], StruNa[16] v1
- jezični alati
  - sustavi za opojavničenje, lematizaciju, MSD-označavanje (Agić–Tadić 2006, Agić–Tadić–Dovedan 2008a, 2008b, 2009, 2010), prepoznavanje imena (*named entity recognition and classificaton, NERC*) (Bekavac 2005), začetci istraživanja parsera (Seljan 2003, Kocijan 2009).

[3] *https://cesar-project.net/.*
[4] *http://filip.ffzg.hr/cgi-bin/run.cgi/first_form.*
[5] *http://riznica.ihjj.hr/.*
[6] *http://nlp.ffzg.hr/resources/corpora/hrwac/.*
[7] *https://hobs.ffzg.unizg.hr.*
[8] *http://nlp.ffzg.hr/resources/corpora/hrenwac/.*
[9] *http://nlp.ffzg.hr/resources/corpora/setimes/.*
[10] *https://hml.ffzg.hr.*
[11] *https://hjp.srce.hr.*
[12] *http://crodip.ffzg.hr/.*
[13] *http://www.hidra.hr/eurovoc/eurovoc1.HTM.*
[14] *http://norma.hidra.hr/rjecnik/.*
[15] *https://crown.ffzg.hr/.*
[16] *https://struna.ihjj.hr/.*

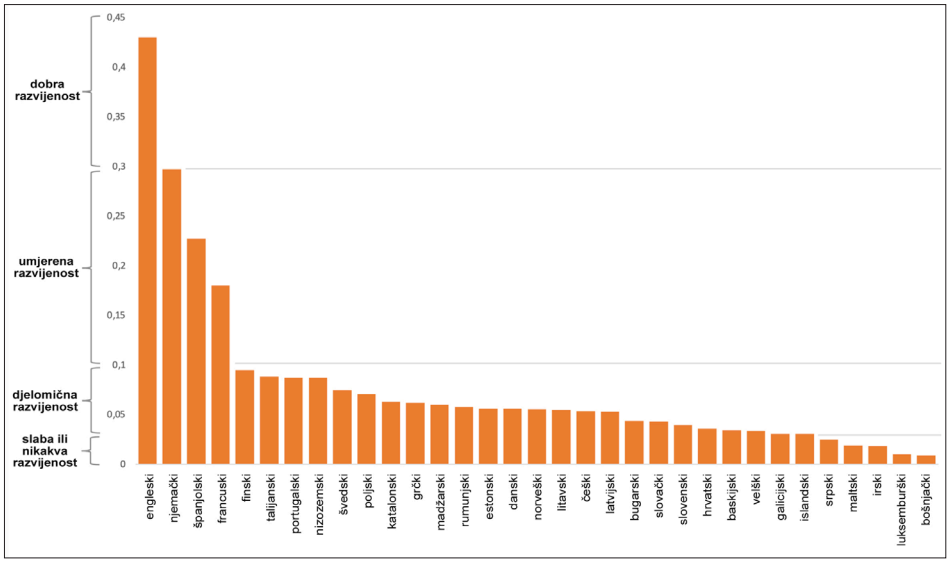
Međutim, tada su postojali i znatni nedostatci u razvoju JT-a za hrvatski jezik koji su bili vidljivi u npr. neusklađenosti formata zapisa jezičnih resursa i međusobnoj neinteroperabilnosti resursa i alata, nedostajali su veliki sintaktički i semantički obilježeni korpusi kao i veliki usporedni korpusi. Premda se hrvatski pojavio u Googleovu prevoditelju već 2008., još uvijek je nedostajalo vlastitih sustava za strojno (potpomognuto) prevođenje. Također nismo imali multimodalnih korpusa tj. govornih korpusa u audio/video zapisu i s transkripcijom, a osobito nam je nedostajala obrada hrvatskoga govora (u oba smjera) kao i dijaloški sustavi. Svi su se ti nedostatci pokušali ukloniti tijekom sljedećih godina istraživanja u području JT-a, a koliko se u tome uspjelo moglo se vidjeti u sljedećoj sličnoj kampanji procjene stanja razvoja JT-a za europske jezike.

3. Stanje razvoja jezičnih tehnologija za hrvatski jezik 2021.

Godine 2021. i 2022. u okviru potpore Europskoga parlamenta, odvijali su se projekti Europska jezična ravnopravnost I. i II.[17] (*European Language Equality, ELE*) u kojima je istražen stupanj razvoja JT-a i jezično usmjerene umjetne inteligencije (UI) za 85 europskih jezika. Ova su dva projekta rezultirala s 35 izvješća o pojedinim jezicima ili skupinama jezika. O razvoju JT-a za hrvatski izvijestilo se u Tadić 2022, a objavljena je i knjiga *European Language Equality* (Rehm–Way 2023) sa sažetim prikazima nalaza u izvješćima za svaki pojedini jezik, a za hrvatski je to objavljeno u Tadić 2023.

Kao što se može uočiti na Sl. 1., hrvatski je jezik tek zakoračio u kategoriju »djelomična razvijenost JT-a« koja je sljedeća nakon najniže kategorije »slaba ili nikakva razvijenost JT-a«. Većina je službenih jezika EU-a u toj istoj kategoriji, a jedino su maltski i irski u nižoj kategoriji. Jedini jezik koji je u kategoriji »dobra razvijenost JT-a« je engleski, a njemački, španjolski i francuski u kategoriji su »umjerena razvijenost JT-a«.

[17] *https://european-language-equality.eu/*.



Slika 1. Stanje razvijenosti JT-a za odabrane europske jezike (Rehm–Way 2023).

Međutim, u razmaku od 10 godina, koliko je proteklo između ova dva istraživanja, ipak je uočljiv napredak u razvoju JT-a za hrvatski jezik ponajprije u većem broju računalnih korpusa i njihovim većim opsezima kao i u većem broju digitalnih rječnika dostupnih ili mrežno ili u aplikacijama za mobilne uređaje. Razvijeni su i alati za obradbu hrvatskih tekstova na sintaktičkoj (prvi ovisnosni parser 2012.) i semantičkoj razini (npr. Wikifier 2014., a u HOBS su 2016. dodane i semantičke uloge). Već je od 2012. hrvatski jezik uključen u sustave EU-a za strojno prevođenje, a 2020. napravljen je prvi domaći strojnoprevoditeljski sustav *Prevoditelj za predsjedanje Vijećem EU-a*, tada za 6 BLEU[18] postotnih bodova bolji od Googleova Prevoditelja za jezične parove englesko-hrvatski i hrvatsko-engleski. Godine 2023. završen je projekt Nacionalna platforma za jezične tehnologije koja je donijela novu inačicu strojnoprevoditeljskoga sustava i postala horizontalna usluga tijelima javne uprave u RH slobodno dostupna na mrežnoj adresi hrvojka.gov.hr. Godine 2016. završen je i projekt prvoga portala za besplatno učenje hrvatskoga kao stranoga jezika (*hr4eu.hr*, *hr4eu.eu*), a još ga uvijek rabi nekoliko tisuća korisnika iz cijeloga svijeta. Veći dio potpore za sve te projekte dolazio je iz EU-a jer se striktno poštivalo jedno od njezinih temeljnih prava, a to je ravnopravnost službenih jezika, pa tako hrvatski kao najmlađi službeni jezik EU-a nije smio zaostajati za jezi-

[18] BLEU je algoritam za automatsko vrjednovanje kakvoće strojnih prijevoda koji su razvili Papineni i dr. (2002), a u načelu se iskazuje kao broj u intervalu između 0 i 1 ili kao postotak u intervalu od 0 do 100. Što je broj viši, to je kakvoća prijevoda viša.

cima koji su to postali prije njega. To je istraživačkim timovima iz Hrvatske otvorilo mogućnost dobivanja financijske potpore i daljnjega razvoja JT-a za hrvatski jezik.

U godini 2021. snimak stanja JT-a za hrvatski jezik pokazivao je sljedeće napretke u ravoju:

- jezični resursi
    - korpusi (v. i na corpora.clarin.hr)
        - ◇ jednojezični: HNK v3[19], hrWaC v2.1[20], *Universal Dependencies* hrvatska banka stabala[21], Hrvatska ovisnosna banka stabala HOBS v2, MARCELL hrvatski pravni korpus[22], čitav niz specijaliziranih korpusa, mrežni korpusi iz velikih kampanja prikupljanja tekstova s mreže (C4Corpus, C100, Deltacorpus, EU Patenti, OSCAR, OPUS...), itd.
        - ◇ višejezični: prijevodi pravne stečevine EU (prije i nakon 2013.)[23], hrenWaC v2[24], MARCELL hrvatsko-engleski usporedni korpus pravnih tekstova[25], ParlaMint v2.1[26], itd.
    - rječnici i leksičke baze: StruNa v2, Školski rječnik[27] i drugi mrežni rječnici i nazivoslovne baze Instituta za hrvatski jezik i Leksikografskoga zavoda Miroslav Krleža[28], CroWN v2.0[29], hrLex v1.3[30], rječnici Školske knjige u aplikaciji za

19 *https://corpora.clarin.hr/crystal/#dashboard?corpname=HNK_v30.*

20 *http://nlp.ffzg.hr/resources/corpora/hrwac/ i http://hdl.handle.net/11356/1064 u repozitoriju CLARIN.SI.*

21 *https://github.com/UniversalDependencies/UD_Croatian-SET/tree/master.*

22 *https://elrc-share.eu/repository/browse/marcell-croatian-legislative-subcorpus/26d822cf443011ea913100155d026706519cacaa114446e4979ae92220adb691/.*

23 *https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en.*

24 *http://hdl.handle.net/11356/1058 u repozitoriju CLARIN.SI.*

25 *https://elrc-share.eu/repository/browse/marcell-croatian-english-parallel-corpus-of-legislative-texts/278f1236443011ea913100155d026706dc6f3bb672cc452c891d08cb9a7bdd69/.*

26 *http://hdl.handle.net/11356/1432 u repozitoriju CLARIN.SI.*

27 *https://rjecnik.hr/.*

28 *http://nazivlje.hr/.*

29 *http://hdl.handle.net/20.500.14615/19 u repozitoriju CLARIN.HR.*

30 *http://hdl.handle.net/11356/1232 u repozitoriju CLARIN.SI.*

mobilne uređaje[31], CroDeriv[32], DerivBase.HR[33], itd.

- jezični alati
  - pojava novih alata za opojavničenje, lematizaciju, MSD-označavanje[34], NERC[35], prvi ovisnosni parser (Agić 2012), uključivanje hrvatskoga u lance obrade jezika (*Language Processing Chain, LPC*) kao što su Udpipe[36], GATE[37], WebLicht[38], CLASSLA Stanza[39], itd.
- proizvodi i usluge
  - Wikifier[40] i EventRegistry[41]: pronalaženje pojmova iz Wikipedije u tekstu, praćenje vijesti o pojedinim događajima (hrvatski uključen uz stotinjak drugih jezika)
  - hrvatski jezik doseže dobru poziciju u semantičkoj mreži Babelnet[42] s oko 3 milijuna sinskupova
  - hrvatski je jezik uključen u uslugu strojnoga prevođenja Europske komisije MT@EC i eTranslation[43] od 2012.
  - strojni Prevoditelj za predsjedanje Vijećem EU-a[44] objavljen je 2020.
  - godine 2023. završava projekt Nacionalna platforma za jezične tehnologije *(National Language Technologies Platform, NLTP)*[45] u kojem je razvijen sustav za strojno (potpomognuto) prevođenje Hrvojka[46] i koji je postao horizontalna usluga za tijela javne uprave u Hrvatskoj

31 *https://apps.apple.com/hr/app/rje%C4%8Dnici/id1076707860?l=hr.*
32 *http://croderiv.ffzg.hr/.*
33 *https://takelab.fer.hr/data/derivbasehr/.*
34 *http://hdl.handle.net/11356/1393 u repozitoriju CLARIN.SI.*
35 *http://hdl.handle.net/11356/1322 u repozitoriju CLARIN.SI.*
36 *https://lindat.mff.cuni.cz/services/udpipe/.*
37 *https://cloud.gate.ac.uk/shopfront/index#tagged=Croatian.*
38 *https://weblicht.sfs.uni- tuebingen.de/weblicht/.*
39 *http://hdl.handle.net/11356/1259 u repozitoriju CLARIN.SI.*
40 *https://wikifier.org.*
41 *https://eventregistry.org.*
42 *https://babelnet.org.*
43 *https://webgate.ec.europa.eu/etranslation.*
44 *https://hr.presidencymt.eu.*
45 *https://www.nltp-info.eu/.*
46 *https://hrvojka.gov.hr/.*

- godine 2016. završava projekt HR4EU s portalom za besplatno učenje hrvatskoga kao stranoga jezika[47] (Filko–Farkaš–Hriberski 2016; Farkaš–Filko–Tadić 2016), a kasnije slične portale razvijaju Matica iseljenika i Croaticum[48] Filozofskoga fakulteta Sveučilišta u Zagrebu.

Međutim, i dalje postoje nedostatci u pojedinim područjima razvoja JT-a, a to se osobito moglo uočiti uslijed nedovoljne količine i veličine velikih usporednih korpusa, nepostojanja multimodalnih korpusa, daljnjega nepostojanja sustava za obradu hrvatskoga govora (u oba smjera) i vrlo malo velikih jezičnih modela (VJM) tj. 2021. se hrvatski jezik pojavio u nekoliko višejezičnih VJM-ova, ali još uvijek nije postojao jednojezični VJM.

4. Promjena paradigme u računalnoj obradi prirodnoga jezika: veliki jezični modeli

Posvemašnja se promjena paradigme u računalnoj obradbi prirodnoga jezika i jezičnim tehnologijama, ponajprije u strojnom prevođenju, počela događati 2014. s uvođenjem postupaka računalnoga dubokoga učenja (*deep learning*) i neuronskih mreža (*neural networks*) s pomoću kojih se mogu izgraditi računalni modeli ljudskoga jezičnoga znanja i jezične uporabe nazvani **veliki jezični modeli** (VJM-i) tj. *Large Language Models* (*LLMs*). Godine 2018. pojavljuju se prvi generativni VJM **GPT-1** (Radford i dr. 2018) i prvi diskriminativni VJM **BERT** (Devlin i dr. 2019), a pravi je pak potres izazvao izlazak VJM-a **ChatGPT** u opću javnost u studenome 2022.

Ovdje bi se moralo posegnuti za bitnim pojmovnim i nazivoslovnim razjašnjenjem. U općoj se javnosti za ChatGPT i slične sustave najčešće koristi nazivak **umjetna inteligencija** (UI). To je neprecizna uporaba nazivka jer je UI znatno šire područje od same obradbe prirodnoga jezika. Naime, ona je samo jedna od sastavnica UI-ja, a ostale su robotika, umjetni vid, pretraga i crpljenje obavijesti, tehnologije znanja, predstavljanje znanja (npr. grafovi znanja), (zdravorazumsko) zaključivanje, itd. Stoga je umjesto »umjetna inteligencija (UI)« za sustave poput ChatGPT-ja i sličnih točniji i precizniji nazivak »veliki jezični model (VJM)«, a takvu ćemo uporabu poštivati i u ovom tekstu.

[47] *https://hr4eu.eu* i *https://hr4eu.hr.*

[48] *https://a1.ffzg.unizg.hr* i *https://a2.ffzg.unizg.hr.*

Odnose između opće UI, generativne UI, VJM-a i **generativnih predobučenih transformatora** (*Generative Pretrained Transformers, GPT*), kao posebne vrste VJM-a, najlakše je predočiti u obliku grafikona (v. Sl. 2).

Slika 2. Shematski prikaz odnosa između opće UI, generativne UI, VJM-a i GPT-a.

Što su veliki jezični modeli? Riječ je o opsežnim jednojezičnim ili višejezičnim skupovima tekstnih podataka koji se koriste za kondenzirano predstavljanje ljudskoga znanja pri uporabi jezika. Opseg tih skupova podataka mjeri se u milijardama tekstova i oni se koriste za prethodnu obuku tj. "trening" VJM-a. Tek obučeni VJM-i mogu "znanje" stečeno tijekom obuke primijeniti na novim tj. do tada neviđenim tekstovima.

Kako obuka izgleda, najlakše je objasniti na primjeru obuke prijevodnih modela. Njih se obučava stalnim stimuliranjem jednom vrstom ulaznoga podatka, a od računalno modelirane mreže (umjetnih "neurona") traži se izlazni podatak što sličniji drugoj polovici ulaznih podataka. Kao primjer navodimo par sravnjenih rečenica između kojih postoji prijevodna ekvivalencija tj. rečenica na ciljnom jeziku prijevod je rečenice s izvornoga jezika.

```
<tu>
      <tuv xml:lang="en">
            <seg>John loves Mary.</seg>
      </tuv>
      <tuv xml:lang="hr">
            <seg>Ivica voli Maricu.</seg>
      </tuv>
</tu>
```

Slika 3. Primjer sravnjenoga para rečenica u odnosu prijevodne ekvivalencije.

Prijevodni se VJM obučava tako da se računalnoj neuronskoj mreži nekoliko milijuna puta daje na ulaznom sloju mreže isti ulazni podatak (u našem primjeru rečenica na engleskom), od unutarnjih se slojeva mreže traži da se veze između "neurona" restrukturiraju u svakom prolazu ulaznoga podatka ne bi li izlazni podatak, koji sustav isporučuje kroz izlazni sloj mreže, bio potpuno jednak ili što sličniji željenom izlaznom podatku (u našem primjeru rečenica ljudski prevedena na hrvatski). Nakon nekoliko milijuna prolaza unutarnji se slojevi mreže rekonfiguriraju kako bi proizveli željeni izlazni podatak što je najviše moguće sličan željenom izlaznom podatku i kad dođe do zasićenja tj. kad se više ne postiže napredak, obuka se prekida i obuka čitavoga VJM-a je okončana. Dodatno se VJM obučen za opći jezik može doobučiti (*fine-tuning*) za različita specijalistička područja (npr. medicina, meteorologija, kemija, pravo, itd.) ili različite zadatke obradbe prirodnoga jezika (prepoznavanje imena ili stavova u tekstovima, crpljenje nazivlja, itd.).

Kako su danas najpoznatiji VJM-i ChatGPT[49] i GPT-5, Gemini[50], Llama[51], Copilot[52], Claude[53], Deep Seek[54], itd. mahom u vlasništvu velikih američkih, europskih ili kineskih tehnoloških tvrtki, s tim je povezano nekoliko ozbiljnih pitanja, a sva se dotiču transparentnosti tih sustava. Naime, najčešće nije poznato koji su se podatci koristili za obuku tih VJM-ova. Nije poznato koliko se često obnavljaju i kakvom vrstom novih tekstnih podataka. Nije poznato uključuje li se u te VJM-e tekstove s posebnim skupovima vrjednota, pa onda oni pokazuju naklonost prema određenim stavovima, a odbojnost prema drugima. Svi su ti VJM-i, na žalost, pušteni na tržište bez ozbiljne ljudske provjere njihove vjerodostojnosti i

49 *https://chatgpt.com/*.
50 *https://gemini.google.com/*.
51 *https://www.llama.com/*.
52 *https://copilot.microsoft.com/*.
53 *https://claude.ai/*.
54 *https://www.deepseek.com/*.

istinitosti. Stoga nisu nepoznate pojave potpuno lažnih tvrdnji (tzv. haluciniranje) koje su iskazane na savršeno kultiviranom jeziku, ali riječ je jednostavno o lažima. Tome se ne treba čuditi. VJM-i su obučeni milijardama tekstova koje su napisali ljudi na temelju svoga jezičnoga znanja i svoje jezične porabe, a sami ljudi u tekstove uključuju štošta. Mnogi govore istinu, ali se u tekstovima može i lagati, izrugivati se, skrivati poruku, mistificirati, zamagljivati, manipulirati sugovornicima, producirati se, itd. Svi su ti tekstovi korišteni za obuku, pa se ne treba čuditi da se takve jezični pojave nalaze i u VJM-ima. Oni su samo odslik onoga što i ljudi čine s jezikom i zapravo ne bismo smjeli tvrditi kako VJM-i imaju vlastitu inteligenciju. Riječ je samo o velikim količinama ljudske uporabe jezika i iz toga automatski izvedenih pravila uporabe na svim jezičnim razinama, pa je stoga s pomoću njih moguće generiranje tekstova na gotovo besprijekornom prirodnom jeziku, ali s istinitim ili posvema neistinitim sadržajem.

Hrvatski je jezik do 2024. bio uključen samo u višejezične (vjVJM-e) jer do tada nije postojao jednojezični (jjVJM) za hrvatski. Riječ je o sljedećim vjVJM-ima:

- Language-agnostic BERT Sentence Encoder (LaBSE): 109 jezika (Feng i dr. 2022)
- X-MOD: 81 jezik (Pfeiffer i dr. 2022)
- CroSloEngualBERT: 3 jezika (Ulčar–Robnik-Šikonja 2020)
- CRO-CoV-cseBERT: dodatno osposobljen za tekstove povezane s COVID-om 19 (Babić i dr. 2021)
- BERTiæ: 4 jezika, zajednički model za bošnjački, hrvatski, crnogorski, srpski (Ljubešić–Lauc 2021)
- XLM-R-parla: 30 jezika (Mochtak i dr. 2023)
- EUBERT: 24 službena jezika EU-a (2023.) [55]
- HR-RoBERTa: prvotno obučen za makedonski jezik, a dodatno hrvatskim podatcima (2021.)[56]
- TwHIN-BERT: preko 100 jezika (Zhang i dr. 2022)

Većina je do sada razvijenih VJM-a višejezična i u njima broj jezika može varirati od dva do više od dvije stotine. To dovodi do neuravnoteženoga skupa podataka (tekstova) koji se koristi za obuku tih VJM-a. Naime, jezici s velikom količinom podataka za obuku dobro su zastupljeni u modelu, a jezici s malom količinom ograničeni su samim nepostojanjem dovoljno podataka za obuku. To dovodi do prezastupljenosti nekih jezi-

[55] *https://huggingface.co/EuropeanParliament/EUBERT.*
[56] *https://huggingface.co/macedonizer/hr-roberta-base.*

ka, najčešće engleskoga, na temelju kojega se uspostavljaju osnovni odnosi u neuronskoj mreži, a potom se ona mijenja ili dopunjuje podatcima iz drugih jezika. Time se zacijelo suprimiraju neke od jezičnih struktura inherentno zastupljene u jezicima s manje podataka za obuku, a umjetno facilitiraju jezične strukture zastupljene u jezicima s velikim količinama podataka. To nesumnjivo dovodi do izobličene reprezentacije "manjih" jezika u vjVJM-ima i njihove podzastupljenosti. Prava bi digitalna jezična ravnopravnost morala pristupiti obuci vjVJM-a tek kad se za svaki uključeni jezik osigura ista količina podataka brojano u npr. pojavnicama.

Kako su VJM-i "crne kutije" i nerijetko se nalaze i u području poslovne tajne tj. ne znamo kako je točno strukturirana njihova neuronska mreža, zbog toga zapravo ne znamo hoće li vjVJM imati bolje performanse kad se primijenjuje na pojedinačni jezik od jednojezičnoga VJM-a (jjVJM) posebno obučenoga samo za taj jedan jezik. Naime, ne možemo bez provjere i vrjednovanja znati hoće li jezici B, C, D, ... Z ometati ili pomagati vjVJM-u kad ga se primijenjuje za jezik A. Kako vrjednovati tu primjenu vjVJM-a na jezik A? Dva su moguća pristupa: 1) izraditi čitav niz posebno prilagođenih alata za vrjednovanje "ponašanja" vjVJM-a za različite jezike, ili 2) izraditi jjVJM za jezik A i onda njegove performanse uzeti kao polaznu vrijednost od koje se može mjeriti otklon u performansama vjVJM-a pri njihovoj uporabi za jezik A. Valja voditi računa kako je pristup 1) još složeniji jer primjena u različitim zadatcima može imati različite performanse. Primjerice, ostali jezici (npr. slovenski, slovački, srpski, bošnjački, crnogorski, češki, ruski, itd.) u vjVJM-u primijenjenom u zadatku ovisnosnoga parsanja hrvatskih rečenica, mogu biti od pomoći tj. mogu pospješiti njegovu učinkovitost u kvaliteti sintaktičke analize hrvatskih rečenica zbog sličnosti sintaktičkih struktura između tih jezika. Međutim, u nekom drugom zadatku, npr. crpljenja nazivaka iz hrvatskih tekstova, prisutnost srpskoga u vjVJM-u unosit će šum i sniziti njegovu učinkovitost za taj zadatak. Naime, riječi, koje su u hrvatskome nazivoslovno obilježene (npr. "pečurka" kao agronomski nazivak tj. prijevod za vrstu gljive koja se na francuskom zove "champignon" ili "crveno vino" kao enološki nazivak za posebnu podvrstu crnoga vina), u srpskome jeziku pak pripadaju općem leksičkom fondu ("pečurka" na srpskom znači svaku 'gljivu' tj. generički je naziv za 'gljive', a "crveno vino" znači 'crno vino' tj. generički je naziv za sva 'crna vina') i ta razlika može dovesti do lošijih performansi vjVJM-a.

K tome, premda je još uvijek prevladavajuće mišljenje kako vjVJM-i postižu bolje rezultate u svakom zadatku računalne obradbe prirodnoga jezika, ipak Martin i dr. (2020) za francuski, Papadimitriou i dr. (2023) za engleski i Virtanen i dr. (2019) za finski pokazuju kako VJM treniran na ne-

koliko jezika (dvoj/trojVJM) ili na jednom jeziku (jjVJM) daje bolje rezultate u nizu zadataka obradbe prirodnoga jezika od vjVJM-a s više desetaka ili stotina jezika.

Prethodna dva razloga bili su poticaj da se tijekom 2024. napravi prvi hrvatski jednojezični VJM, pa je tako u okviru EU-projekta HR-XR-XTEND[57] razvijen HR-GPT Beta (Štefanec i dr. 2024; Thakkar i dr. 2024) slobodno dostupan u hrvatskom repozitoriju jezičnih podataka europske istraživačke infrastrukure CLARIN[58]. Taj je hrvatski jjVJM obučen na 7,75 milijardi riječi hrvatskih tekstova i prerast će u punu inačicu kad ga se obuči s ukupno 14 milijardi riječi koje za sad postoje prikupljene, ali još čekaju potrebnu predobradbu.

5. Razvoj jezičnih tehnologija za hrvatski jezik: trajna zadaća

Razvoj jezičnih tehnologija za pojedini jezik nesumnjivo omogućuje jeziku da ostane živ i uporabiv u digitalnome okružju, a osobito u komunikacijskim kanalima 21. stoljeća. Ma kakvi god oni postali u budućnosti, nije za očekivati da ćemo kao vrsta *Homo sapiens* uskoro razviti telepatiju, pa ćemo zacijelo i dalje morati komunicirati na nekom prirodnom jeziku. Ako za hrvatski jezik ne budemo imali razvijene jezične tehnologije, koje će olakšati njegovu uporabu u tim komunikacijskim kanalima, onda će ljudi, koji se njima koriste, iz čiste komocije posegnuti za nekim jezikom za koji takve tehnologije postoje. Toga će trenutka hrvatski postati funkcionalno nepismen ili "digitalno nepismen" jezik jer se postojanje jezičnih tehnologija za neki jezik danas može smatrati novim stupnjem pismenosti. Takav će jezik ostati onkraj digitalne razdjelnice i ubrzo će mu prijetiti izumiranje jer će se, do tada njegovi, govornici početi koristiti nekim drugim jezikom i postati govornici nekoga drugoga jezika.

Iz prethodno rečenoga sasvim je razvidno kako će se jezici, koji ne budu imali svoje jednojezične VJM-e ili ne budu sudjelovali u najpopularnijim višejezičnim VJM-ima, naći isključeni iz novih kretanja u razvoju jezičnih tehnologija, a takvo što nikako ne bismo poželjeli hrvatskome jeziku. Zbog toga kao jezična zajednica moramo osigurati dovoljno jezično i sadržajno kvalitetnoga, općega i specijaliziranoga jezičnoga gradiva na hrvatskome jeziku tj. podataka za obuku VJM-a kako bi se 1) mogao obučiti jednojezični hrvatski VJM za svaku novu arhitekturu VJM-a koja će se pojaviti i 2)

[57] *https://hr-xr-xtend.ffzg.unizg.hr/*.

[58] *http://hdl.handle.net/20.500.14615/2-4* u repozitoriju CLARIN.HR: *https://repository.clarin.hr.*

hrvatski jezik uvijek pojavio u svakom važnijem višejezičnom VJM-u zasebno, a ne pod kapom nekakvoga HBS-makro-jezika kao što se u nekima od projekata izrade VJM-a već dogodilo.

Vitalnost hrvatskoga jezika i jamstvo njegova preživljenja u digitalnome okružju i umjetnoj inteligenciji u budućnosti možemo osigurati samo njegovom stalnom prisutnošću u podatcima za obuku raznih VJM-ova u dovoljnoj količini. Stoga valja organizirati i prionuti na kontinuiranu široku kampanju prikupljanja do sada nezabilježene količine tekstnih podataka na hrvatskome jeziku i ponuditi ih istraživačima za izradu VJM-a. Imajući to u vidu, Zakon o hrvatskom jeziku doista jest zakon za budućnost jer se u njegovu članku 16, stavak 2c jasno regulira kako se u Nacionalnom planu hrvatske jezične politike svakako mora razraditi unaprjeđenje jezičnih tehnologija za hrvatski jezik. Štoviše, navode se neka od područja JT-a od posebnoga značaja, a jedno je i »...slobodan pristup digitaliziranim tiskovinama na hrvatskom jeziku te drugi javni mrežni jezični servisi...« koji bi neizostavno morali biti u trajnom otvorenom pristupu i na usluzi svima. Naime, VJM-i su danas temelj za budući razvoj JT-a, ali i za razvoj tehnologija znanja koje se na VJM-e nadograđuju. Umjetna inteligencija mora znati hrvatski jer to njegovi govornici već sad traže.

Bibliografske referencije

Agić, Željko. 2012. *Pristupi ovisnosnom parsanju hrvatskih tekstova*, doktorska disertacija, Sveučilište u Zagrebu, Filozofski fakultet.

Agić, Željko; Marko Tadić. 2006. Evaluating Morphosyntactic Tagging of Croatian Texts. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genova, European Language Resources Association (ELRA). *http://www.lrec-conf.org/proceedings/lrec2006/pdf/326_pdf.pdf*.

Agić, Željko; Marko Tadić; Zdravko Dovedan. 2008a. Combining Part-of-Speech Tagger and Inflectional Lexicon for Croatian. Ur. Erjavec, T.; J. Žganec Gros. *Proceedings of the 6th Language Technologies Conference*. Ljubljana, Institut Jožef Stefan, 116–121.

Agić, Željko; Marko Tadić; Zdravko Dovedan. 2008b. Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. *Informatica (Ljubljana)*, 32 (4), 445–451.

Agić, Željko; Marko Tadić; Zdravko Dovedan. 2009. Evaluating Full Lemmatization of Croatian Texts. Ur. Klopotek, M.; Przepiorkowski, A.; Wierzchon, S. & Trojanowski, K. *Recent Advances in Intelligent Information Systems*. Varšava: Academic Publishing House EXIT, 175–184.

Agić, Željko; Marko Tadić; Zdravko Dovedan. 2010. Tagger Voting Improves Morphosyntactic Tagging Accuracy on Croatian Texts. Ur. Lužar-Stiffler; V., Jarec, I. & Bekić, Z. (ur.) *Proceedings of the 32nd International Conference on Information Technology Interfaces*. Zagreb, SRCE University Computer Centre, University of Zagreb, 61–66.

Babić, Karlo; Milan Petrović; Slobodan Beliga; Sanda Martinčić-Ipšić; Mihaela Matešić; Ana Meštrović. 2021. Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model. *Applied Sciences*, 11(21), 10442. MDPI AG. *http://dx.doi.org/10.3390/app112110442* (pristupljeno 25. 3. 2025.).

Bekavac, Božo. 2005. *Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima*. Doktorska disertacija. Zagreb: Filozofski fakultet.

Boras, Damir. 2005. *Znanstveni projekt 130464 Hrvatska rječnička baština i prikaz rječničkoga znanja, Popis digitaliziranih rječnika*. Sveučilište u Zagrebu, Filozofski fakultet. *http://crodip.ffzg.hr/rjecnici.pdf* (pristupljeno 25. 3. 2025.).

Devlin, Jacob; Ming-Wei Chang; Kenton Lee; Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis: ACL, 4171–4186. *https://aclanthology.org/N19-1423/* (pristupljeno 30. 3. 2025.).

Hrvatski opći leksikon. 2012. mrežno izdanje. Zagreb: Leksikografski zavod Miroslav Krleža. *https://hol2.lzmk.hr/clanak/tehnologija* (pristupljeno 25. 3. 2025.).

Kocijan, Kristina. 2009. *Model parsera za hrvatski jezik*. Doktorska disertacija. Zagreb: Filozofski fakultet.

Farkaš, Daša; Matea Filko; Marko Tadić. 2016. Hr4eu – using language resources in computer aided language learning. *Proceedings of the Second International Conference Computational Linguistics in Bulgaria*. Sofia: The Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences, 38–44.

Feng, Fangxiaoyu; Yinfei Yang; Daniel Cer; Naveen Arivazhagan; Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin: ACL, 878–891.

Filko, Matea; Daša Farkaš; Diana Hriberski. 2016. Hr4eu – a Web-portal for e-Learning of Croatian. Ur. Salomi Papadima-Sophocleous; Linda Bradley; Sylvie Thouësny. *CALL communities and culture – short papers from EUROCALL 2016*. Limassol: Research-publishing.net, 137–143. *https://files.eric.ed. gov/fulltext/ED572005.pdf* (pristupljeno 25. 3. 2025.).

László, Bulcsú; Svetozar Petrović. 1959. Uvod. Ur. Bulcsú László; Svetozar Petrović. Strojno prevođenje i statistika u jeziku. *Naše teme* 3(6), 105–298.

Ljubešić, Nikola; Davor Lauc. 2021. BERTić – the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. Ur. Bogdan Babych i dr. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Kiyv: ACL, 37–42.

Martin, Louis; Benjamin Muller; Pedro Javier Ortiz Suárez; Yoann Dupont; Laurent Romary; Éric de la Clergerie; Djamé Seddah; Benoît Sagot. 2020. CamemBERT: a tasty French language model. Ur. Dan Jurafsky i dr. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 7203–7219. *https://aclanthology.org/2020.acl-main.645* (pristupljeno 25. 3. 2025.).

Mochtak, Michal; Peter Rupnik; Nikola Ljubešić. 2023. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. arXiv preprint arXiv:2309.09783. *https://arxiv.org/abs/2309.09783* (pristupljeno 25. 3. 2025.).

Papadimitriou, Isabel; Kezia Lopez; Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. Ur. Lisa Beinborn i dr. *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, Dubrovnik: ACL, 143–146.

Papineni, Kishore; Salim Roukos; Todd Ward; Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. U: *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL2002)*, 311–318. *https://aclanthology.org/P02-1040.pdf* (pristupljeno 24. 3. 2025.).

Pfeiffer, Jonas; Naman Goyal; Xi Victoria Lin; Xian Li; James Cross; Sebastian Riedel; Mikel Artetxe. 2022. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. Ur. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle: ACL, 3479–3495.

Radford, Alec; Karthik Narasimhan; Tim Salimans; Ilya Sutskever. 2018. *Improving language understanding by generative pre-training*. OpenAI. *https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf* (pristupljeno 25. 3. 2025.).

Rehm, Georg; Hans Uszkoreit (ur.) 2012. *META-NET White Paper Series: Europe's Languages in the Digital Age*, Heidelberg: Springer. *https://european-language-equality.eu/meta-net-white-paper-series/list-of-volumes/* (pristupljeno 25. 3. 2025.).

Rehm, Georg; Andy Way (ur.) 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cham: Springer. *https://link.springer.com/content/pdf/10.1007/978-3-031-28819-7.pdf?pdf=button* (pristupljeno 25. 3. 2025.).

Seljan, Sanja. 2003. *Leksičko-funkcionalna gramatika hrvatskoga jezika: teorijski i praktični modeli*. Doktorska disertacija. Zagreb: Filozofski fakultet.

Skansi, Sandro; Leo Mršić; Ines Skelac. 2020. A Lost Croatian Cybernetic Machine Translation Program. Ur. Sandro Skansi. *Guide to Deep Learning Basics*. Cham: Springer, 67–78.

Štefanec, Vanja; Daša Farkaš; Gaurish Thakkar; Marko Tadić. 2024. Building a Large Language Model for Croatian. Ur. Constantin Oraşan; Tharindu Ranasinghe; Gloria Corpas Pastor; Ruslan Mitkov; Maria Kunilovskaya; Vilelmini Sosoni; Marie Escrib. *Proceedings of the Conference New Trends in Translation and Technology (NeTTT2024)*. Varna: Incoma, 204–209. *https://acl-bg.org/proceedings/2024/NeTTT%202024/pdf/2024.nettt-1.17.pdf* (pristupljeno 25. 3. 2025.).

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex libris.

Tadić, Marko. 2009. New Version of the Croatian National Corpus. Ur. Dana Hlaváčkova i dr. *After Half a Century of Slavonic Natural Language Processing*. Brno: Masarykovo sveučilište, 199–205.

Tadić, Marko; Dunja Brozović-Rončević; Amir Kapetanović. 2012. *Hrvatski jezik u digitalnom dobu / The Croatian Language in the Digital Age*. Heidelberg: Springer, 2012.

Tadić, Marko. 2022. *D1.7 Report on the Croatian Language*, izvješće u projektu European Language Equality. *https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_7__Language_ Report_Croatian_.pdf* (pristupljeno 25. 3. 2025.).

Tadić, Marko. 2023. Language Report Croatian. Ur. Georg Rehm; Andy Way. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cham: Springer, 111–114.

Thakkar, Gaurish; Vanja Štefanec; Daša Farkaš; Marko Tadić. 2024. Building a Large Language Model for Moderately Resourced Language: A Case of Croatian. Ur. Neven Vrček; Dina Korent. *Proceedings of the 35th Central European Conference on Information and Intelligent Systems*. Varaždin: Sveučilište u Zagrebu, Fakultet organizacije i informatike, 225–232. *https://archive.ceciis.foi.hr/public/conferences/2024/Proceedings/S6/S6P3.pdf* (pristupljeno 25. 3. 2025.).

Ulčar, Matej; Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. *Proceedings of the Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*. Brno: Springer, 104–111.

Virtanen, Antti; Jenna Kanerva; Rami Ilo; Jouni Luoma; Juhani Luotolahti; Tapio Salakoski; Filip Ginter; Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv 2019, arXiv:1912.07076. *https://arxiv.org/abs/1912.07076* (pristupljeno 25. 3. 2025.).

Zhang, Xinyang; Yury Malkov; Omar Florez; Serim Park; Brian McWilliams; Jiawei Han; Ahmed El-Kishky. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. arXiv Preprint arXiv:2209.07562. *https://arxiv.org/abs/2209.07562* (pristupljeno 25. 3. 2024.).

Why are language technologies important for the future of the Croatian language?

Abstract

The paper provides a definition, composition and overview of Language Technologies (LT) as well as the results of two large-scale evaluation campaigns of the development status of LT for dozens of European languages in which Croatian participated. While in the META-NET campaign in 2011, Croatian was placed in a group of languages with underdeveloped language technologies, in the European Language Equality campaign in 2021, it was placed in a group of twenty languages with partially developed language technologies and relative progress was shown. However, it was in the development of language technologies that large language models (LLMs) introduced a paradigm shift and showed that a large number of language tools developed so far will have to be reproduced with a new methodology in the background, i.e. one that uses LLMs. The paper further provides an overview of the basic types of LLMs and explains the difference between artificial intelligence and LLMs. The paper is concluded by pointing out the need for the continuous development of LT for the Croatian language, based precisely on the continuous development of new LLMs for the Croatian language in accordance with any new LM-architecture that will emerge. This requires the provision of unprecedented quantities of textual data in Croatian. Otherwise, the Croatian language may experience "digital illiteracy" and remain beyond the digital divide.

Ključne riječi: jezične tehnologije, hrvatski jezik, veliki jezični modeli

Keywords: Language Technologies, Croatian language, Large Language Models